ABSTRACT

Title of proposal: Gathering Language Data Using Experts

Denis Peskov, 2022

Dissertation directed by: Professor Jordan Boyd-Graber
Department of Computer Science
College of Information Studies
Language Science Center
Institute for Advanced Computer Studies

Natural language processing needs substantial data to make robust predictions. Automatic methods, unspecialized crowds, and domain experts can be used to collect conversational and question answering NLP datasets. A hybrid solution of combining domain experts with the crowd generates large-scale, free-form language data.

A low-cost, high-output approach to data creation is **automation**. We create and analyze a large-scale audio question answering dataset through text-to-speech technology. Additionally, we create synthetic data from templates to identify limitations in machine translation. We conclude that the cost-savings and scalability of automation come at the cost of data quality and naturalness.

Human input can provide this degree of naturalness, but is limited in scale. Hence, large-scale data collection is frequently done through **crowd-sourcing**. A question-rewriting task, in which a long information-gathering conversation is used as source material for many stand-alone questions, shows the limitation of using this methodology for **generating** data. Certain users provide low-quality rewrites—

removing words from the question, copy and pasting the answer into the question—if left unsupervised. We automatically prevent unsatisfactory submissions with an interface, but the quality control process requires manually reviewing 5,000 questions.

Therefore, we posit that using domain **experts** for data generation can create novel and reliable NLP datasets. First, we introduce computational adaptation, which adapts, rather than translates, entities across cultures. We work with native speakers in two countries to generate the data, since the gold label for this is subjective and paramount. Furthermore, we hire professional translators to assess our data. Last, in a study on the game of Diplomacy, community members generate a corpus of 17,000 messages that are self-annotated while playing a game about trust and deception. The language is varied in length, tone, vocabulary, punctuation, and even emojis. Additionally, we create a real-time self-annotation system that annotates deception in a manner not possible through crowd-sourced or automatic methods. The extra effort in data collection will hopefully ensure the longevity of these datasets and galvanize other novel NLP ideas.

However, experts are expensive and limited in number. **Hybrid** solutions pair potentially unreliable and unverified users in the crowd with experts. We work with Amazon customer service agents to generate and annotate of goal-oriented 81,000 conversations across six domains. Grounding the conversation with a reliable conversationalist—the Amazon agent—creates free-form conversations; using the crowd scales these to the size needed for neural networks.

Gathering Natural Language Processing Data Using Experts

by

Denis Peskov

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Jordan Boyd-Graber, Chair
Professor Philip Resnik, Dean's Representative
Professor Michelle Mazurek
Professor Katie Shilton
Professor John Dickerson

Dedication

The spirit is willing, but the flesh is weak. $\rightarrow$ *Russian* $\rightarrow$ ...

The vodka is strong, but the meat is rotten.

-Georgetown-IBM, 1950

The spirit is willing, but the flesh is weak.

-Google Translate, 2021

The spirit desires, but the flesh is weak.

-Yandex Translate, 2021

The spirit is willing, but the flesh is weak.

-DeepL Translate, 2021

If as one people speaking the same language they have begun to do this, then nothing they plan to do will be impossible for them.

Acknowledgments

Thank you to my advisor Jordan Boyd-Graber; my committee members Philip Resnik, Michelle Mazurek, Katie Shilton, and John Dickerson; Joe Barrow and peers from the University of Maryland CLIP lab; Alex Fraser, Hinrich Schütze, and others from Ludwig-Maximilians-Universität München; Benny Cheng, Sander Schulhoff, and other students I've supervised; ARLIS, the DAAD, Amazon Research, 3M, Raytheon BBN, and other sources of funding or collaboration; Tom Hurst and the Computer Science Department; the Language Science Center; Tim Beach, Terrence Reynolds, and other undergraduate professors that inspired me; my College Park roommate Kodjo Aflagah; my family; and my friends.

I could not have done this without you.

Table of Contents

List of Tables

List of Figures

Chapter 1: The Case for Upfront Investment in Data

Computation can solve tasks across multiple areas of scientific inquiry: natural language processing, computer vision, biology. Solving tasks for each of these domains—translating a sentence between languages, distinguishing a cat from a dog, classifying a mutation—has two abstract and intertwined dependencies: model-building and data collection.[1] The relationship is intertwined since today's models are optimized to draw statistical conclusions from significant amounts of data through machine learning. But, even the most cutting edge modeling techniques are heavily dependent on having *realistic* and *accurate* data for solving a task. These large datasets are primarily gathered from online repositories or created through low-cost crowd-sourcing (Deng et al., 2009; Rajpurkar et al., 2016; Budzianowski et al., 2018), which are often *artificial* or *inaccurate*. We argue that high-quality, **expert**-reliant data collection can lead to long-term improvements in Natural Language Processing (NLP) and enable complex, novel tasks.

[1] Mitchell (1997) defines a machine learning model as, "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E". This E depends on data collection.

1.1 Defining Data: Annotation and Generation

In the overview, we discuss the two tasks necessary for data collection and explain the importance of data quality for computer science as a field.

Data creation can be broadly categorized into two categories: **generation** and **annotation**. We define **generation** as the creation of a data item—sequencing a genome, creating a new image, gathering a new sentence from a user, or automatically creating a sentence (Atkins et al., 1992; Goodfellow et al., 2014; Zhu et al., 2018)—that is not previously **found** elsewhere (Section 2.2.1). We define **annotation** as the application of a label to an existing data item (e.g., classifying a part of the genome, labeling an image as a cat, or describing the sentiment of a sentence) (Deng et al., 2009; Finin et al., 2010; Kozomara and Griffiths-Jones, 2014). In many fields, data must be both **generated** to be representative of the task and then accurately **annotated** to be effective.

1.2 Quantity over Quality as a Paradigm

The demand of neural models for quantity has caused models to be trained on large, noisy data (Brown et al., 2020). The building blocks of other research areas—gene sequences in biology and individual pixels in computer vision—are not readily human interpretable by default. Even in more human-intuitive fields, like natural language processing, data have reached the scale where their veracity—the certainty and completeness of the data—cannot be assumed (Qiu et al., 2016),

despite the early assertions by Atkins et al. (1992). They posit that, "there is in fact little danger of obfuscation for the major parameters that characterize a corpus: its size (in numbers of running words), and gross characterizations of its content."[2] However, the objectivity of size is questionable; a corpus consisting of the same word repeated a million times clearly differs from one with a million unique words. Yet size remains a primary consideration.

This focus on quantitative metrics evaluation metric has shaped NLP data creation during the past decade (Rodriguez et al., 2021). A dataset paper will comment on the amount of words, sentences, questions, etc., but with no assessment of their quality. But, the sheer quantity of data masks biases and artifacts, as they are no longer obvious to the naked human eye (Pruim et al., 2015; Gururangan et al., 2018; Gor et al., 2021a). Since current approaches to machine learning often obscure how decisions are made by a model, the quality of the data is not immediately questioned as a culprit when a false prediction is made.

The current paradigm of crowd-sourcing—"the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call"' (Howe et al., 2006)—for dataset creation has been the main impetus of unreliability in data. Specifically, natural language processing has generally depended on low-cost crowds following an exploration of crowd-sourcing for five NLP tasks (Snow et al., 2008) and ImageNet (Deng et al., 2009). However, ImageNet's entirely crowd-sourced annotations still have notable problems after a decade of updates (Yang et al., 2020)

[2]Additionally, they crucially comment that the evaluation of corpora has not been standardized.

and should serve as a cautionary tale. Incorrect annotations of a cat in place of a dog may be a trivial mistake in a stand-alone context, but other unchecked errors—intentional or even hostile—can percolate into significant real world problems: racist suggestions and hate speech chatbots (Mac, 2021; Hunt, 2016). Hence, some tasks benefit from large-scale data, some require high quality data, and some require both quality and quantity. A re-prioritization to working with users that have a reputation incentive to generate realistic and reliable data would benefit tasks in the latter two categories.

1.3 The Nuance of Using Text as Data

We introduce the Natural Language Processing tasks covered in our work, challenges faced in NLP due to trade-offs of annotation speed and quality, and the distinction between generation and annotation.

A large focus of NLP is on building models that exploit patterns in language data to solve a variety of tasks: question answering, conversational agents, machine translation, information extraction, etc. Our research builds three models—logistic regression (Section 2.3.1), deep averaging networks (Section 2.3.3) and long short-term memory networks (Section 2.3.4)—for two tasks—question answering (Section 2.1.2) and dialog (Section 2.1.3). However, in the current paradigm of machine learning, models answer questions or predict dialog acts based on existing training data. This makes *realistic* data a prerequisite for any model that aims to *realistically* solve a language task.

But, the prevalence of neural models in NLP has prioritized data size over realism. Chapter 2 describes the relationship between data and natural language processing tasks. At the extreme end, GPT-3 is trained on 499 *billion* tokens, *de facto* training a neural model based on the entire Internet (Brown et al., 2020). However, not everything on the Internet is relevant or accurate! This is significant since training data containing low-quality data unsurprisingly leads to models learning controversial or false conclusions, with high levels of confidence (Wolf et al., 2017; Wallace et al., 2019a). Therefore, missing or false data in the data **generation** process undermines the ability of NLP to realistically solve language tasks.

Furthermore, many tasks in NLP depend on accurate **annotation** of the raw data. As a thought experiment, if all verbs are labeled as nouns and all nouns are labeled as verbs in the training data, a perfectly designed language model would be confidently wrong in its predictions. Crowd-sourcing with generalists (Buhrmester et al., 2011) assumes that enough unspecialized workers will answer a question correctly. This is a valid assumption for unambiguous, multiple-choice annotation with a large amount pool of annotators. However, many annotation tasks, such as span-annotation or candidate selection, have so many parameters that they are akin to language **generation**, which cannot be easily verified through IAA (Karpinska et al., 2021). The annotation can be error prone due to the amount of items that need annotation: discourse may require annotating ellipses or co-referential links that could be easily missed by different users (Orăsan, 2003). Or each item requiring annotation may have thousands of options, such as an open-ended question (i.e., answering "Who was a famous mathematician?" from a potential candidate pool of

all named entities in Wikipedia). Therefore, NLP **annotation** needs to be accurate, at least in aggregate.

1.4 Data Quality as a New Paradigm

Investing in reliable data—as defined by its **generation** and **annotation** dimensions—upfront has two benefits. First, this improvement in the quality and diversity of data is a prudent long-term investment as high-quality datasets can have shelf-lives of decades (Marcus et al., 1993; Miller, 1995a) while model architectures are frequently supplanted (Vaswani et al., 2017; Peters et al., 2018; Devlin et al., 2019a). Second, using experts for data generation can enable tasks not otherwise possible; generalists cannot annotate medical images nor generate sentences in a language which they do not speak.

We use experts in three experiments to collect NLP corpora and contrast them with past automated and crowd-sourced ones. First, we show the limitations of using automated methods of data collection (Chapters 3 and 4). Second, we show that crowd-sourcing can **generate** data flexibly but inaccurately (Chapter 5). Third, we show the merits of using experts as **annotators** for data evaluation for a subjective and novel named entity adaptation task (Chapter 6). Fourth, we describe an experiment that uses experts for both **generation** and **annotation** to study deception through the medium of a board-game (Chapter 7). Last, we discuss a hybrid approach—using verified experts paired with external, low-cost data sources (Vukovic and Bartolini, 2010) (Chapter 8) that can mitigate some of the

Chapter	Data Source	Data Type	Task
3	Automation	Generation	Question Answering
4	Automation	Generation	Dialog
5	Crowd-Sourced	Generation	Question Answering
6	Expert	Annotation	Question Answering
7	Expert	Generation, Annotation	Dialog
8	Hybrid	Generation, Annotation	Dialog

Table 1.1: A tabular summary of our projects. Our thesis is organized in increasing order of data source complexity.

accuracy issues while scaling in size and cost. We classify our work by data source, data type, and applicable task (Table 1.1).

Chapter 2: Natural Language Processing Depends on Data

This chapter discusses several NLP tasks that require data and the types of data collection discussed in this thesis.

Each NLP task requires its own bespoke training data, such as text in multiple languages for machine translation. Certain tasks within the subfields of question answering and dialog (Section 2.1) are unable to be solved with naturally-**found** data and require dataset creation.

Different types of users can **generate** and **annotate** the data needed for these language models. Unspecialized users can be asked to solve tasks through **crowd-sourcing** and automated methods can generate data at scale (Section 2.2.3). **Experts** can gather and annotate data (Section 2.2.4). Last, **hybrid** approaches combine anonymous crowd users with experts that verify the results (Section 2.2.5). We provide the necessary background and past work relevant to these three user types (Section 2.2). We explain the models and metrics that are used in solving these tasks (Section 2.3).

2.1 Tasks

Language models can be created for different NLP tasks, but each requires a different type of training data. We focus on two NLP tasks in our research: Question Answering and Dialog.

2.1.1 What is a Task?

According to Resnik (2022), a task is an abstraction connected with a real-world problem that enables us to compare potential solutions in order to assess progress and consists of:

1. the real-world problem you care about
2. the dataset that's going to be used
3. the definitions of input/output that methods will need to use
4. the measurements that will be used to quantify performance
5. the criteria for what constitutes "better", i.e. progress

2.1.2 Question Answering

Question answering (QA) is one task heavily dependent on training data. The five-fold task tuple (Section 2.1.1) for QA is:

1. **real world problem:** information retrieval
2. **data:** datasets such as those found in Table 6.3
3. **input/output**: input of free form text and an output of a selected span from

Questions
What is the English meaning of caliente?
What is the meaning of caliente (in English)?
What is the English translation for the word “caliente”?

Table 2.1: Three questions from TREC 2000 data that are believably varied. The test questions were carefully crafted by experts.

existing text or a free form answer

4. **evaluation:** answer accuracy, perhaps weighted by question difficulty
5. **standard for progress:** higher accuracy and the ability to answer more complicated types of questions

In the current machine learning paradigm, QA generally answers a question with a previously seen answer. Therefore, the coverage of questions and answers is important as models trained on trivia questions are unlikely to answer inquiries about medical symptoms, and vice versa. We discuss the relevant history of question answering and review the most relevant datasets.

The Text Retrieval Conference established QA as an annual, formalized task (Voorhees et al., 1999). The questions were carefully curated every year and modifications to the question answering task were made. Table 2.1 shows examples of questions that are intended to fool systems reliant on literal information extraction.

Machine reading comprehension, “a task introduced to test the degree to which a machine can understand natural languages by asking the machine to answer questions based on a given context” (Li et al., 2019), ushered in larger more diverse QA datasets, with SQuAD (Rajpurkar et al., 2016, 2018) being the most popular leaderboard for models. The amount of questions went from being measured in the

Questions	Answers
"Which laws faced significant opposition?"	later laws
"What was the name of the 1937 treaty?"	Bald Eagle Protection Act

Table 2.2: The paper examples from SQuAD. Unlike Table 2.1, these questions are done through crowd-sourcing and Wikipedia and are not carefully planned.

hundreds to being measured in the *hundreds of thousands*. Example questions are provided in Table 2.2. Large influential question answering datasets include SQuAD 1.0 (Rajpurkar et al., 2016), SQuAD 2.0 (Rajpurkar et al., 2018), MS Marco (Bajaj et al., 2016), TriviaQA (Joshi et al., 2017) QuAC (Choi et al., 2018), Quizbowl (Rodriguez et al., 2019), and Natural Questions (Kwiatkowski et al., 2019). We summarize the size of these datasets and their user pools in Table 6.3.

Computers can read a question and select the answer from a passage of text. This format of QA is called machine reading comprehension (Rajpurkar et al., 2016, MRC), and has been a popular choice for dataset design. However, QA models struggle to generalize when questions do not look like the standalone questions systems in training data: e.g., new genres, languages, or closely-related tasks (Yogatama et al., 2019). Unlike MRC, **conversational question answering** requires models to link questions together to resolve the conversational dependencies between them: each question needs to be understood in the conversation context. For example, the question *"What was he like in that episode?"* cannot be understood without knowing what *"he"* and *"that episode"* refer to, which can be resolved using the conversation context. CoQA is a conversational question answering dataset addressing different domains—Wikipedia, children's stories, news articles, Reddit, literature, and science articles—created by pairing Mechanical Turk crowd-sourced workers

Dataset	# of Questions	Data Source
CoQA	8,000	Crowd
SQuAD 1.0	100k	Crowd
SQuAD 2.0	50k	Crowd
QuAC	100k	Crowd
TriviaQA	95k	Hybrid
Quizbowl	100k	Hybrid
Natural Questions	300k	Hybrid
MS Marco	1000k	Found
TREC-8	200	Expert
Trick Me	651	Expert

Table 2.3: A tabular summary of dialog datasets. The datasets described as hybrid all scrape or use naturally-occurring language and then supplement it with crowd-sourced annotation.

together (Reddy et al., 2019).

Recent work acknowledges that certain community practices around crowd-sourcing may not be optimal for QA. Boyd-Graber (2020) question the paradigm of using crowd-sourced workers as the measure for human baselines, rather than evaluating through a play test. As one alternative to the crowd, Wallace et al. (2019b) work with the Quizbowl community to rewrite questions to be adversarial (and evaluate with a play test). At the intersection of question answering and machine translation, Clark et al. (2020a) emphasize that natural speakers of a language must be used to write authentic questions in languages outside of English.[1]

2.1.3 Dialog

The five-fold task tuple (Section 2.1.1) for dialog is:

1. **real world problem:** automating conversation

[1]Although the source of these speakers is still crowd-sourced unverified users as they do not have other scalable access to speakers of typologically diverse languages.

2. **data:** datasets such as those found in Table 2.4
3. **input/output**: input and output of free form text
4. **evaluation:** naturalness of response or completion of end goal
5. **standard for progress:** comparisons to human dialogs such as the Turing Test (Turing, 1950), in which a computer tries to fool a human into thinking its a human through textual communication

Existing **found** conversational data has been repurposed as NLP datasets. Ubuntu threads provide millions of conversations of technical support (Lowe et al., 2015). Reddit, a collection of threaded comments about diverse subjects, and OpenSubtitles, collections of movie and television subtitles, provide millions of sentences as training data (Henderson et al., 2019).

However, **found** datasets cannot cover all domains and languages. For example, the audio data needed to automatically generate subtitles are unlikely to exist in low-resource languages, customer service data for training a chat bot is proprietary, and defendants are unlikely to carefully annotate sentences where they are lying in a court deposition (nor is the court likely to release the court deposition in a machine readable format). Therefore, **generating** conversational datasets becomes a NLP need. The Dialog State Tracking Challenge (Henderson et al., 2014) creates several relatively-small, crowd-sourced datasets focusing on different conversational areas on an annual basis. MultiWOZ proposes a framework for simulated conversations, which is necessary for domains containing sensitive data that cannot be released (Budzianowski et al., 2018).

Dataset	# of Questions	Data Source
DSTC2	1,612	Found
Ubuntu Dialog	930,000	Found
Reddit	256,000,000	Found
OpenSubtitles	316,000,000	Found
DSTC2	1,612	Crowd
CoQA	8,000	Crowd
MultiWOZ	8,438	Crowd

Table 2.4: A tabular summary of key dialog datasets.

2.2 Data Collection Type

Data for machine learning can come from one of five sources: finding data, automation, crowd-sourcing, experts, and a hybrid combination thereof. We discuss representative works for each of these data pools.

2.2.1 Finding

Reusing existing text through scraping websites or forums and re-purposing historical documents can create datasets with little effort. We define the this type of data as **found**.

The Internet contains troves of data, but this data is noisy due to having a low barrier to entry for contributors. Amazon reviews (McAuley et al., 2015), Twitter (Banda et al., 2021), and Wikipedia (Vrandečić and Krötzsch, 2014) provide language from aliased and often anonymous users.

In contrast, organizations that have an incentive to control or report their data release accurate, or at least authentic, datasets. EuroParl is collected from professionally translated official parliamentary proceedings (Koehn, 2005). Literature

comes from a verified author (Iyyer et al., 2016), as do Reuters news articles (Lewis et al., 2004). The United Nations maintains detailed datasets about global populations. The World Trade Organization releases a comprehensive collection of legal disputes. Enron released authentic emails sent by verifiable employees when its problems spilled out into the public domain (Klimt and Yang, 2004).

What is the common denominator of these datasets? This data is sourced from **experts** (e.g., World Trade Organization lawyers and translators) or unverified online users (e.g., Reddit users). However, since this data was not originally created for NLP, further data processing and **annotation** are often required.

2.2.2 Automation

Data **generation** is necessary as the data necessary for NLP cannot always be found. Synthetic data can be created according to fixed rules or templates, which we refer to as automation. Augmentation is a frequent phrasing of this way of creating data (Kafle et al., 2017). This method can create datasets of any scale, but it does not guarantee their authenticity.

Templates can create datasets unlimited in scale, but dubious in realism. Filatova et al. (2006) generate questions using specific verbs for various domains: airplane crashes, earthquakes, presidential elections, terrorist attacks. In their own words, their automatically created templates are "not easily readable by human annotators" and the evaluation requires a lengthy discussion. Examples of questions generated though templates include the following ambiguous questions about *specific*

earthquakes:

- *Is it near a fault line?*
- *Is it near volcanoes?*

Chapter 3 describes our project in which text-to-speech creates a dataset of 500,000 audio files. While large, our dataset is limited to a single female voice and read in a notably different cadence than that of realistic Quizbowl experts. Additionally, our automation method depends on the existence expert-written questions in the first place. However, to create a dataset of the same size with human experts would require thousands of hours. Mozafari et al. (2014) propose using active learning to minimize the human effort needed to gather large-scale datasets; one gathers annotations for a subset of the data and then extrapolates those labels to similar unlabeled data. This serves as a segue into the next type of data creation method: crowd-sourcing.

2.2.3 Crowd-Sourcing

We define crowd-sourcing techniques, explain their history, and comment on the repercussions of the wide-spread use of this data pool in NLP today. Crowd-sourcing is, 'the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call" (Howe et al., 2006). Crowd-sourcing, in the applied sense, relies on unspecialized users and is the most popular way to create new annotated datasets in NLP today.

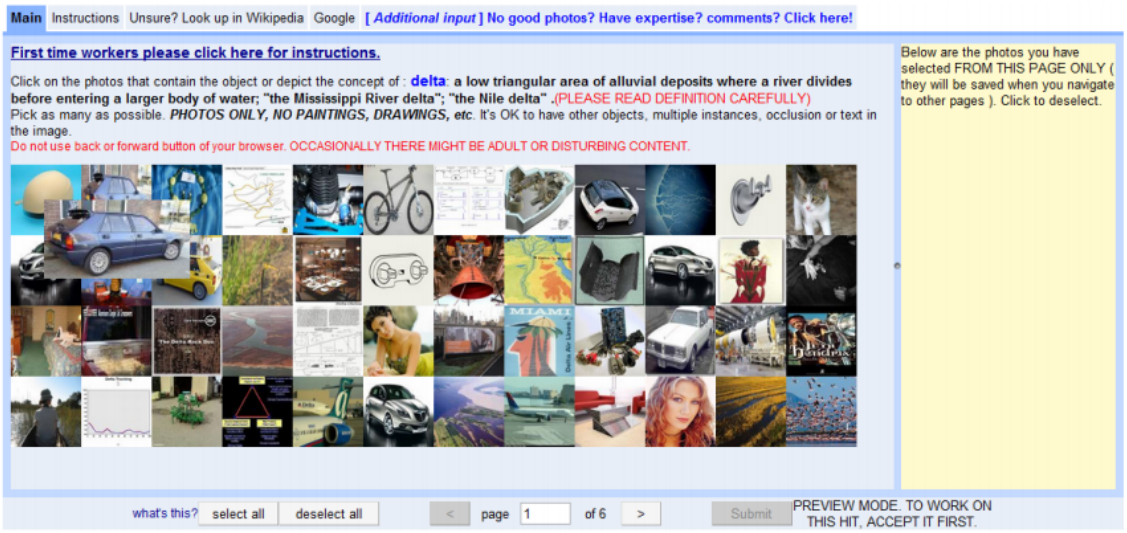

Figure 2.1: Deng et al. (2009) popularizes Mechanical Turk use for Computer Science. Simple **annotation** tasks can be completed reliably with crowd-sourcing since selecting if an image belongs to a WordNet category (e.g., car, bicycle, delta) is a relatively objective and straightforward task, despite the occasional gray area (i.e., the image contains both a car *and* a bicycle or the bicycle is in a cubist painting). However, many NLP tasks are not so clear-cut as a different contemporary study showed (Snow et al., 2008).

ImageNet became an influential work in computer science that used crowd-sourcing for cheap annotation. Deng et al. (2009) build ImageNet by crowd-sourcing image annotations for WordNet. Visual classification tasks are maximally simple in nature since annotators are asked to decide if an image contains a Burmese cat. Figure 2.1 shows their interface.

Despite their effort to simplify and explain the task, disagreement is a major problem and a minimum of *10* users are used to guarantee a level of confidence. Even with constant updates, the dataset still has limitations a decade later from the initial scaling methodology used to create it (Yang et al., 2020). Would training and rewarding the annotators upfront have saved time and money in the long-run?

Crowd-sourcing spread to disciplines other than machine vision as a source

for research data. Mechanical Turk, the platform used for ImageNet, became the largest crowd-sourcing marketplace by making it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can complete these tasks virtually (Amazon, 2021). Buhrmester et al. (2011) claim that Amazon Mechanical Turk gathers "high-quality data inexpensively and rapidly" for psychology. The average psychology experiment is conducted using university students that require hourly compensation and usually come from a concentrated geographic area and socio-economic background. However, the evidence for this claim stems from having participants fill out a survey and is primarily evaluated on the time required, rather than the quality of the final result. In their survey, users report that their motivation for using Mechanical Turk is higher on a Likert scale for enjoyment than for payment. Given that nearly every NLP task requires that users complete a large amount of previous tasks (1000+) and with a nearly perfect accuracy (90%+), this claim seems unlikely to hold for the average producer of NLP data. As a note of caution, Mason and Suri (2012) claim that spammers are likely to target surveys on Mechanical Turk.

Crowd Flower, renamed as Figure Eight, is a platform similar to Mechanical Turk, but with a focus on quality control. While Mechanical Turk keeps track of **Human Intelligence Tasks** (HIT)—the name for each individual task—accuracy rates, this metric depends on task providers to manually evaluate the data and provide feedback about the worker. This level of oversight is unlikely to occur for thousands of tasks. Crowd Flower's innovation is to include a test set with each task which monitors that users' responses correspond to gold labels. As early

adopters of crowd-sourcing, Finin et al. (2010) use Crowd Flower for annotating named entities in Twitter. However, most annotations are completed by a few prolific workers, which opens up the dataset to potential biases. Furthermore, creating a crowd-sourced dataset with Crowd Flower is possible for **annotation** but not for **generation**.

The success of computer vision annotations led researches to use crowd-sourcing for collecting annotations in natural language processing tasks such as word sense disambiguation and machine translation (Callison-Burch et al., 2015). Snow et al. (2008) demonstrate that (on average) four non-expert workers can emulate an expert for five NLP tasks: affect recognition, word similarity, textual entailment, temporal event recognition, and word sense disambiguation. Using a nonprofessional user pool is the default manner for collecting large datasets for NLP as it can generated and annotated quickly and cheaply. As one example, large question answering datasets involving Wikipedia and search engines—SQuAD, SearchQA—use crowd-sourcing to generate questions (Rajpurkar et al., 2016; Dunn et al., 2017).

The two main benefits to this data source are the cost and the rapid rate of data collection. The cost is unquestionably lower for an employer or researcher to use the crowd rather than internal employees. Crowd workers are paid a fraction of what full-time employees would receive for the same task and do not receive any benefits (Whiting et al., 2019).[2] Largely due to the variations in cost-of-living around the world and flexibility of the work, the pay is appealing to some workers. The demographics of the platform more accurately model the United States than the

[2]This clearly is not a pro from the worker's perspective.

average college student, at least for psychology experiments (Buhrmester et al., 2011; Difallah et al., 2018). For perspective, Amazon Mechanical Turk has over a hundred-thousand workers, thousands of which are available at any moment (Difallah et al., 2018). Modular tasks can be completed in hours in crowd-sourcing, as thousands of temporary workers complete tasks faster than a handful of employees.

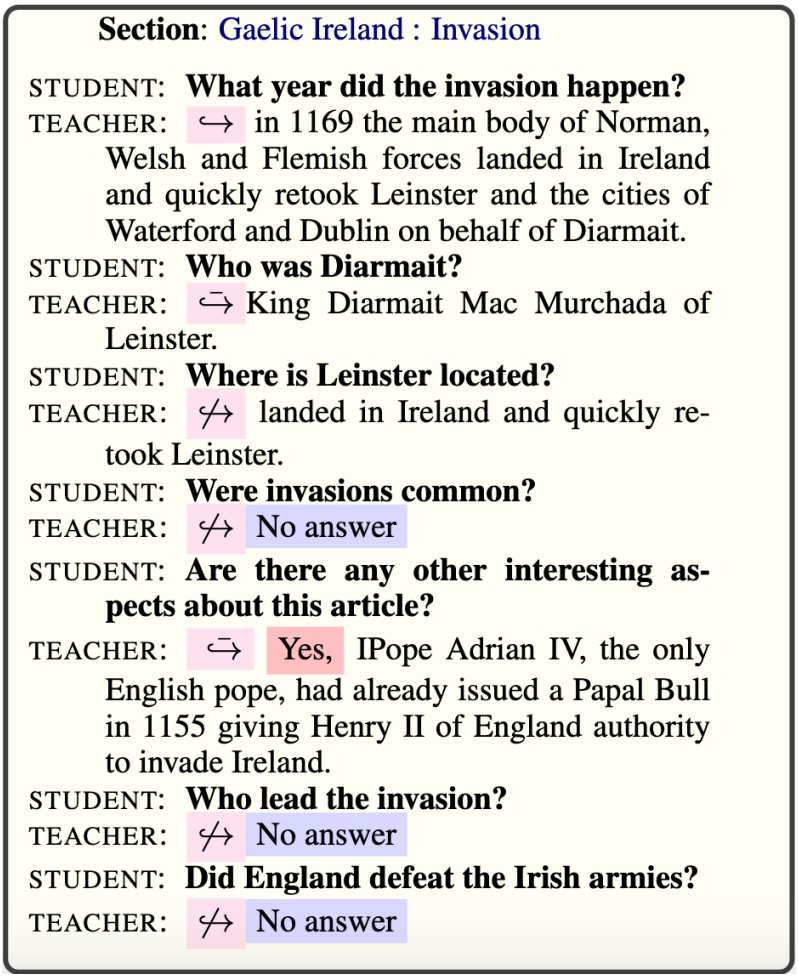

Figure 2.2: Crowd-sourcing can also be used to generate large-scale NLP data. However, **generation** creates a quality issue not present in **annotation**. In this particular example, Choi et al. (2018) highlight that the teacher does not provide quality responses. However, the student's conversation is quite unnatural and has grammatical issues.

The con to crowd-sourcing is that quality control becomes the central chal-

lenge for crowd-sourcing NLP data. Zaidan and Callison-Burch (2011) show that data gathered from crowd-sourcing for machine translation nets a BLEU score nearly half the size of professional translators, and only one point higher than an automatic machine translation approach. Other studies have shown that users tend to voluntarily provide inaccurate data (Suri et al., 2011) and misrepresent their background (Chandler and Paolacci, 2017; Sharpe Wessling et al., 2017).[3] Last, there is an upper-bound to the complexity of crowd-sourced tasks. Crowd workers become less reliable and efficient for tasks that are not straightforward (Finnerty et al., 2013). Figure 2.2 shows that more complicated NLP task instructions are not followed in good faith. For classification tasks, average accuracy needs to exceed 50% for reliable annotators to overcome their noisy peers (Kumar and Lease, 2011). This is not a threshold that is always achievable since answers for certain tasks are sparse.

Chapter 3 reveals quality issues in this technique through a project that crowd-sources question. We use the crowd to rewrite sequential questions into a standalone format. However, extensive manual review is necessary to remove the low-quality contributions from the data pool. Experts are accountable in ways the crowd-user is not and do not require the same level of post-collection quality control.

Chapter 6 uses a crowd-sourced project, WikiData, for its modeling. WikiData (Vrandečić and Krötzsch, 2014) is a structured, human-annotated representation of Wikipedia entities that is actively developed. The method proves less

[3]As a tangential consideration, legal regulation may ultimately limit the effectiveness of this technique, since it is completely unregulated by current employment practices (Wolfson and Lease, 2011).

Message	Sender's intention	Receiver's perception
If I were lying to you, I'd smile and say "that sounds great." I'm honest with you because I sincerely thought of us as partners.	Lie	Truth
You agreed to warn me of unexpected moves, then didn't . . . You've revealed things to England without my permission, and then made up a story about it after the fact!	Truth	Truth
. . . I have a reputation in this hobby for being sincere. Not being duplicitous. It has always served me well. . . . If you don't want to work with me, then I can understand that . . .	Lie	Truth
(Germany attacks Italy)		
Well this game just got less fun	Truth	Truth
For you, maybe	Truth	Truth

Table 2.5: In contrast to the previous conversations involving crowd workers, conversations involving experts **generate** creative, and even humorous, language. Additionally, the **annotation** of truthfulness is not possible with crowd-sourcing, since it requires the **generator's** real-time knowledge. This conversation snippet is from the Diplomacy project (Chapter 7).

accurate than expected due to the entities being unevenly populated.

2.2.4 Expert

We define "experts", provide a brief summary of relevant datasets, and introduce a dataset **generated** and **annotated** by domain experts. We use the definition from Weinstein (1993):

> *"An individual is an expert in the 'performative' sense if and only if he or she is able to perform a skill well."*

Defining expertise is a tricky and subjective goal; for example, "well" is highly subjective in this definition. Bourne et al. (2014) conclude that psychology is the appropriate framework for evaluating expertise, which "results from practice and

experience, built on a foundation of talent, or innate ability". For NLP, we require that the person has both the incentive and skill to *accurately*, as opposed to quickly, complete their task. A degree of accountability, rather than full anonymity, is important as it prevents intentional fraud (Teitcher et al., 2015). Therefore, we require that experts be identifiable, in at least some capacity during the data collection process. Such experts can be trained or they can be found in specialized communities of interest.

The social sciences have traditionally relied on a limited amount of trained experts for establishing quantitative support for a theory. Grounded theory is "the discovery of theory from data" (Glaser and Strauss, 2017). Annotation is called **coding** and is used to systematically categorize and then analyze content (Neuendorf, 2017; Krippendorff, 2018). Baumer et al. (2017) find that the, "grounded theory analysis took two researchers several hours of work per week over roughly two and a half months... With grounded theory, every single response was read and reread multiple times." This level of commitment does not easily scale.

Unsurprisingly, the amount of datasets created with this attention to detail for NLP are limited due to the high cost associated with hiring experts and quality assurance. As an alternative, skilled citizen scientists may generate high-quality language in the pursuit of a hobby such as journalism, writing, or debate (Silvertown, 2009; Rymes and Leone, 2014). Given the increasing investment and interest in the field, this route for data collection will be the best long-term investment. We discuss existing sources of this kind of data, methods for generating language data, and methods for annotating language data.

Language recorded *naturally* for other purposes has led to datasets that have withstood the test of time. The United Nations, New York City, and the World Trade Organization are all organizations that release reliable large-scale data, as discussed in Section 2.2.1. These organizations hire professionals such as translators and lawyers to generate language.

However, existing, or **found**, data sources do not cover all NLP tasks and domains. Therefore, **generation** by experts is necessary. Examples of this include adversarial questions written by trivia players (Wallace et al., 2019b), Document Understanding Conference summaries (Over, 2003), code summaries written by developers (Badihi and Heydarnoori, 2017), and story generation (Akoury et al., 2020).

Annotations are possible to collect from non-experts, but often at the expense of their accuracy. Programmers can self-annotate their code for easier future accessibility (Shira and Lease, 2011). Hate speech annotation is more accurate with expert annotators than amateur ones (Waseem, 2016). In the security field, privacy policies are complicated to understand (and to annotate) for the lay user (Jensen and Potts, 2004; Audich et al., 2018). In the medical field, the lack of expert annotation poses a barrier to large-scale NLP clinical solutions (Chapman et al., 2011). Unsurprisingly, doctor annotation is more accurate than online generalist annotation for medical diagnoses (Cheng et al., 2015).

The quality of crowd-sourced work relative to expert work has been disputed in multiple studies. Mollick and Nanda (2016) compare expert to crowd judgment for the funding of theater productions. They conclude that most decisions are aligned between the two pools, but that crowds are more swayed by superficial presentation

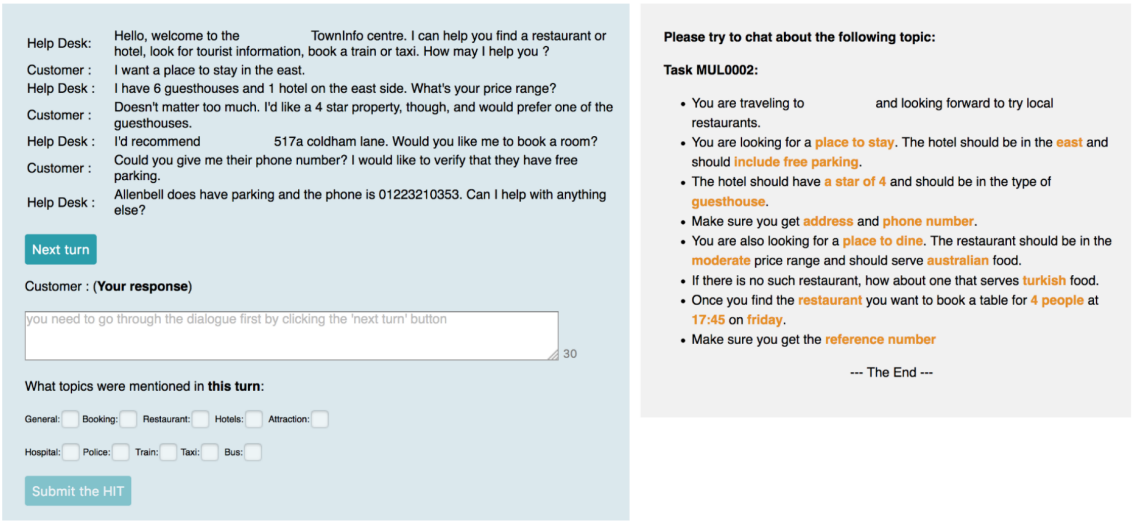

Figure 2.3: Hybrid approaches try to control the quality of language **generated** by the crowd. MultiWoz (Budzianowski et al., 2018), creates a rigid template for the user conversation, avoiding the worst quality issues at the expense of user creativity.

than underlying quality. Leroy and Endicott (2012) compare annotations of text difficulty between a medical librarian and a non-expert user and do not see a large difference on a small sample size.

Chapter 7 presents a project that works with the Diplomacy, a popular board-game, community to **generate** and **annotate** a natural conversational dataset for the task of deception. The language in this dataset is realistic and impossible to generate with unspecialized crowd users. An example conversation is provided in Table 7.1.

2.2.5 Hybrid

Hybrid approaches aim to enhance crowd-sourcing by overseeing unspecialized labor or automatic methods with expert knowledge. This combination lowers cost and allows for data scaling, while maintaining a certain level of quality control.

We define hybrid user pools and discuss past projects.

We define hybrid data collection sources as any that combine a cost-saving pool, such as crowd-sourcing or automation, with expert supervision. This is a natural extension of crowd-sourcing and does not require as detailed of a historical overview: once quality issues were noted, attempts were made to remedy them. For **generation**, crowd-sourced workers can be combined with trained agents to create data for a given NLP task. For **annotation**, crowd-sourced workers can be supervised by trained experts.[4]

As an illustrative example, Zaidan and Callison-Burch (2011) propose an oracle-based approach to identify the high quality crowd-sourced workers and rely on their judgments. The paper claims that crowd-sourcing can lead to a notable reduction in cost without a complete loss in quality. Their approach crucially depends on having expert (professional) translations as a reference point.

Hybrid approaches improve quantity and quality for other NLP tasks. Kochhar et al. (2010) use a hierarchical system for database, specifically Freebase, slot filling. First, an item is populated by automatic methods, then issues are escalated to volunteer users, and any remaining issues are escalated to trained experts. Ade-Ibijola et al. (2012) design a system for essay-grading that allows for teacher oversight and compare their results to area experts. Hong et al. (2018) optimize the productivity of medical field experts by providing additional reference resources and standardizing databases. FEVER (Thorne et al., 2018a) relies on super-annotators on one percent of the data as a comparison point for all other annotations for FEVER. Er-

[4]Automation can replace the crowd for simple tasks in this hybrid approach.

rors made by crowd-sourced workers on Named Entity Recognition can be clustered and identified, which in turn can be escalated to a skilled arbitrator to improve task guidance (Nguyen et al., 2019). Having an expert-written template that crowd workers must follow eliminates the worst-quality submissions (Budzianowski et al., 2018). This example is provided in Figure 2.3. A combination of trained and untrained workers can be used for generating Wizard-of-Oz personal assistant dialog (Byrne et al., 2019).

Furthermore, some crowd-sourcing platforms rely on this hybrid approach. Crowd Flower, mentioned in Section 2.2.3, attempts to bolster the reliability the crowd by requiring the task master to create gold-standard test questions, which are interspersed among the data being collected (Vakharia and Lease, 2015). While not necessarily using experts, this provides an automatic quality filter that down-weights the reliability of annotations made by the least accurate–as determined by the gold-standard test set—annotators. Crucially, this approach can only work for **annotation**, as generation quality cannot be quickly assessed. ODesk is a crowd-sourcing platform that provides a hybrid approach, as it relies on crowd-sourcing from the Internet, but vets the participants to have a matching skill-set for the task (Vakharia and Lease, 2015). Prolific and Upwork are two other platforms that place additional emphasis on vetting reliable users.[5][6]

[5] `http://www.prolific.co`
[6] `http://www.upwork.com`

2.3 Models & Metrics

Data is a prerequisite for machine learning. We summarize popular models that can trained from data. Additionally, we discuss the metrics used to evaluate these models. This emphasis on the model, and not the underlying data, evaluation is a limitation in NLP.

2.3.1 Logistic Regression

According to Ng and Jordan (2002), the **logistic regression** is a basic *discriminative* model, meaning that it can classify items into one of several classes. It relies on using features x to predict class y by learning a vector of weights, $\vec{w}$, and a bias term, b according to:

$$z = \vec{w} \cdot \vec{x} + b \tag{2.1}$$

The variable z is then passed through a sigmoid function to transform the values to a probability:

$$\sigma(z) = \frac{1}{(1 + e^{-z})} \tag{2.2}$$

Additionally, the **loss function** tells the logistic regression how quantitatively wrong a prediction is. Popular loss functions include Cross Entropy Loss—often used for logistic regression and classification tasks— and Mean Squared Error (Sammut and Webb, 2010).

There are two phases to logistic regression: training and testing. During

training, stochastic gradient descent and cross-entropy loss learn the optimal weights of $\vec{w}$ and b. Cross-entropy loss calculates the difference between the predicted $\hat{y}$ and the true y. The gradient descent algorithm (Bottou, 2010; Ruder, 2016) finds the minimum loss.

At test time, for each example the highest probability label is predicted in y. Multinomial logistic regression allows for the prediction of more than two classes.

The logistic regression model is interpretable since the weight of each feature is transparent in the final prediction. Certain features have higher weights than other ones. A feature weight of close to zero would indicate that the feature is not essential for the model; conversely the highest weighted feature is important in the task. This has made the logistic regression a popular baseline model for machine learning as its straightforward interpretability contrasts with the current state-of-the-art model: neural networks.

2.3.2 Neural Models

Neural networks are a more powerful classifier than logistic regressions and can be shown to learn any function due to a **hidden layer**. The hidden layer is a layer that applies a, usually, nonlinear transformation to an input to generate a new output. As a result they often avoid dependence on carefully crafted features and learn their own representations for the task (Jurafsky and Martin, 2000).

All neural networks depend on **backpropagation**. The **hidden layer(s)** allows nonlinear transformations but needs to be trained to produce a desirable

output. This is done through **backpropagation**, which percolates weight adjustment with the chain rule throughout the entire network. The gradient of the loss function is calculated one layer at a time, iterating backwards from the last layer (hence *back*propagation).

We focus on neural architectures applicable to NLP: Deep Averaging Networks (Section 2.3.3) and Recurrent Neural Networks (Section 2.3.4).

2.3.3 Deep Averaging Network

The **Deep Averaging Network**, or DAN, classifier proposes a simple architecture with comparable results to more complicated neural models. Unlike Logistic Regression, the DAN adapts to linguistic versatility by using embeddings in lieu of specific word features. It has three sections: a "neural-bag-of-word" (NBOW) encoder, which composes all the words in the document into a single vector by averaging the word vectors; a series of hidden transformations, which give the network depth and allow it to amplify small distinctions between composed documents; and a softmax predictor that outputs a class.

The encoded representation $\mathbf{r}$ is the averaged embeddings of input words. The word vectors exist in an embedding matrix $\mathbf{E}$, from which we can look up a specific word w with $\mathbf{E}[w]$. The length of the document is N. To compute the composed representation r, the DAN averages all of the word embeddings:

$$\mathbf{r} = \frac{\sum_i^N \mathbf{E}[w_{\mathrm{i}}]}{N} \tag{2.3}$$

The network weights $\mathbf{W}$, consist of a weight-bias pair for each layer of transformations ($\mathbf{W^{(h_i)}}$, $\mathbf{b^{(h_i)}}$) for each layer i in the list of layers L. To compute the hidden representations for each layer, the DAN linearly transforms the input and then applies a nonlinearity: $\mathbf{h_0} = \sigma(\mathbf{W^{(h_0)}r} + \mathbf{b^{(h_0)}})$. Successive hidden representations h_i are: $\mathbf{h_i} = \sigma(\mathbf{W^{(h_i)}h_{i-1}} + \mathbf{b^{(h_i)}})$. The final layer in the DAN is a softmax output: $\mathbf{o} = \text{softmax}(\mathbf{W^{(o)}h_L} + \mathbf{b^{(o)}})$. This model is used and modified in Chapter 3.

2.3.4 Sequence Models

Unlike the DAN, **Recurrent Neural Networks** (Elman, 1990, RNN) take into account the sequence of the input, which is important given the ordered nature of language. The **long short-term memory** (Gers et al., 2000, LSTM) modifies the RNN by allowing it to discard past information.

According to Goldberg (2017), **Sequence to Sequence** refers to a model that ingests a sequence of text and then generates a sequence of text, rather than a single classification, as an output. The architecture necessary for this is called Encoder-Decoder, as the text input is first encoded—meaning a sequence of text has been transformed into a numerical representation—and then decoded—this representation is then transformed back into text.

Machine translation is a clear example where Sequence to Sequence applies. If a sentence in German needs to be transformed into English, then the German sentence is first encoded into a numerical representation and then decoded into an English sentence.

Attention (Bahdanau et al., 2015) is a modification of the LSTM that looks at different parts of the encoded sequence at each stage in the decoding process. Visualizing attention provides a mild level of interpetability as the model looks at a specific part of the input. We use these models in Chapters 7 and 8, as the current state of the art for NLP.

Additionally, rather than relying on n-gram language models, neural language models reference prior context as **embeddings** that represent the word(s). This means that the neural network can understand that "cat" and "dog" are similar, and can be treated similarly, whereas a n-gram model assumes independence. word2vec (Mikolov et al., 2013b) and GloVe (Pennington et al., 2014) embeddings are commonly used pre-trained embeddings. This powerful innovation precipitated the current state-of-the-art dependence on Transformers, which are used in Chapters 7 and 8.

The Transformer model simplifies the architecture and dispenses with recurrences and convolutions (Vaswani et al., 2017), relying instead entirely on attention. ELMo (Peters et al., 2018), used in Chapter 8, improves on GloVe embeddings (Pennington et al., 2014) by allowing a word's embedding to adjust to the context, rather than being committed to having a single word sense. BERT improves the embeddings further by looking at context bidirectionally, meaning that words that follow a word influence its embedding. These pre-trained embeddings can be further fine-tuned to accommodate a specific domain's context.

2.3.5 Evaluation

But how does one evaluate a model, or the underlying quality of data? Model evaluation is specific to a task: classifying images correctly for ImageNet or answering a question for SQuAD. There is a goal of achieving the highest quantitative accuracy on a particular task (Wang et al., 2019a); qualitative analysis of *what* was answered correctly in contrast to another model is an after-thought (Linzen, 2020).

Data evaluation is necessary for crowd-sourcing. For annotation, one can compare the annotations of users to one another using **Inter-Annotator Agreement** (IAA) (Artstein and Poesio, 2008), the most popular of which is Cohen's Kappa (Cohen, 1960). Nowak and Rüger (2010) show that for simple image classification tasks, the majority vote of unspecialized users is comparable to expert annotation. Passonneau and Carpenter (2014) confirm these results comparing trained undergraduates with the crowd for word sense annotation. Additionally, having a large amount of annotators allows them to establish a confidence in the label accuracy for each individual word.

However, there is no obvious metric to compute IAA for **generation**. Machine translation uses metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and TERp (Snover et al., 2009) as an automatic approximation of *target* quality; however, the quality of the *source* data—which must be generated by human users—is never evaluated. In question answering, one may limit the possible answers to existing pages in Wikipedia, or some other finite source, to avoid string matching problems. But, language is complex and multiple users could write equally

valid questions that do not appear similar at the character level. Table 2.1 is one such example. Deng et al. (2021) propose a unified set of metrics for compression, transduction, and creation tasks as a first step in systematically assessing language generation quality.

The pivot to language models, and later the neural revolution in natural language processing precipitated an ever-increasing race for data; the largest dataset, not the best model architecture may be the key differentiating factor for solving a NLP task. But how to evaluate the influence of data rather than architecture is an open research question. Since this is a broad question, we focus on two areas of NLP that are data dependent: question answering and dialog. Four possible sources of data are presented and compared: **found**/**automatic** (Chapters 3 and 4), **crowd-sourced** (Chapter 5), **expert-sourced** (Chapters 6 and 7), and **hybrid** (Chapter 8). A large-scale data project explores the limitations of relying on model accuracy without data verification (Chapter 3).

Chapter 3: Automatic Data Generation from a Found Source[1]

The fastest method of creating large neural-scale datasets is through automatic generation of synthetic data. This chapter discusses a large audio dataset created with Text-To-Speech for the task of question answering (Section 2.1.2), and its limitations (Section 3.1). The dataset, while large, is not realistic and would be supplanted by a similar human-**generated** dataset (Chapter 5). Furthermore, both datasets ultimately depend on **experts** for validation (Chapters 6 and 7).

3.1 Automated Data Creation for Question Answering

Progress on question answering (QA) has claimed human-level accuracy. However, most factoid QA models are trained and evaluated on clean text input, which becomes noisy when questions are spoken due to automatic speech recognition (ASR) errors. This consideration is disregarded in trivia match-ups between machines and humans: IBM Watson (Ferrucci, 2010) on Jeopardy! and Quizbowl (QB) matches between machines and trivia masters (Boyd-Graber et al., 2018) provide text data for machines while humans listen. An Artificial Intelligence needs to process speech

[1]Denis Peskov, Joe Barrow, Pedro Rodriguez, Graham Neubig, and Jordan Boyd-Graber. 2019. Mitigating noisy inputs for question answering. In Conference of the International Speech Communication Association. Peskov is responsible for the data creation, the gathering of recordings from users, running the neural models, figure and table design, and paper writing.

input, akin to how a typical human would process sound, to play the trivia game without this outside assistance.[2]

Unlike a typical human, the computer needs a model to decode the audio into text and answer the question. Unfortunately, there are no large *spoken* corpora of factoid questions with which to train models; text-to-speech software can be used as a method for generating training data at scale for question answering models (Section 5.1). Although synthetic data is less realistic than true human-spoken questions it easier and cheaper to collect at scale, which is important for training. These synthetic data are still useful; models trained on synthetic data are applied to human spoken data from QB tournaments and Jeopardy! (Section 3.4.1).

Noisy ASR is particularly challenging for QA systems (Figure 3.1). While humans and computers might know the title of a "revenge novel centering on Edmund Dantes by Alexandre Dumas", transcription errors may mean deciphering "novel centering on edmond dance by alexander $<unk>$" instead. Dantes and Dumas are low-frequency words in the English language and hence likely to be misinterpreted by a generic ASR model; however, they are particularly important for answering the question. Additionally, the introduction of distracting words (e.g., "dance") causes QA models to make errors (Jia and Liang, 2017). Key terms like named entities are often missing, which is detrimental for QA (Section 3.2.1).

Previous approaches to mitigate ASR noise for answering mobile queries (Mishra

[2]An audibly impaired person would be delivered questions in a non-audio medium, but would still experience a perceptual delay, unlike a machine. The end goal of a trivia QA robot should be considered: a trivia robot that wins every match by buzzing at superhuman speeds does not provide an optimal trivia game for humans. Therefore, robots should be tested on their knowledge, not their response time, unless the goal is to test different robots against one another.

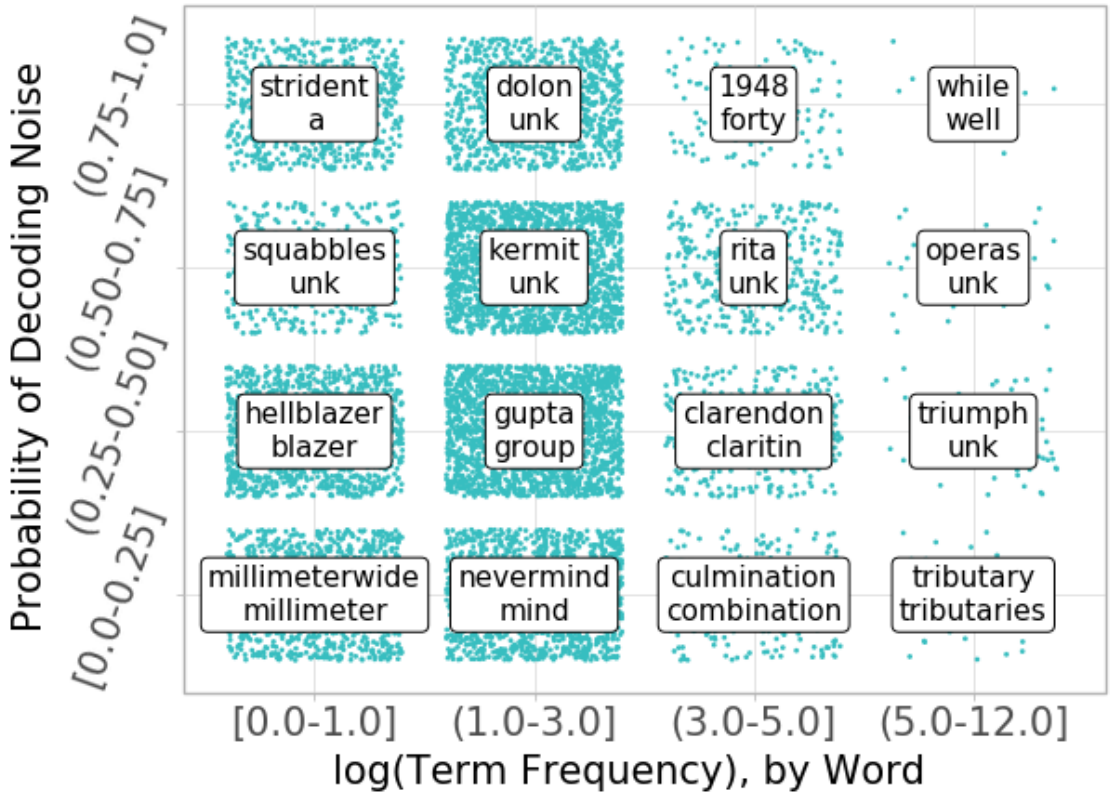

Figure 3.1: ASR errors on QA data: original spoken words (top of box) are garbled (bottom). While many words become into "noise"—frequent words or the unknown token—consistent errors (e.g., "clarendon" to "clarintin") can help downstream systems. Additionally, words reduced to $<unk>$ (e.g., "kermit") can be useful through forced decoding into the closest incorrect word (e.g., "hermit" or even "car").

and Bangalore, 2010) or building bots (Leuski et al., 2009) typically use unsupervised methods, such as term-based information retrieval. Our datasets for training and evaluation can produce *supervised* systems that directly answer spoken questions. Machine translation for speech (Sperber et al., 2017) also uses ASR confidences; we evaluate similar methods on QA.

Specifically, some accuracy loss from noisy inputs can be mitigated through a combination of forcing unknown words to be decoded as the closest option (Section 3.3.2), and incorporating the uncertainties of the ASR model directly in neural

models (Section 3.3.3). The forced decoding method reconstructs missing terms by using terms audibly similar to the transcribed input. Word-level confidence scores incorporate uncertainty from the ASR system into a Deep Averaging Network (Section 2.3.3). These methods are compared against baseline methods on our synthetic and human speech datasets for Jeopardy! and QB (Section 3.4).

3.2 Automatically Generating a Speech Dataset

Neural networks require a large training corpus, but recording hundreds of thousands of questions is not feasible. Methods for collecting large scale audio data include Generative Adversarial Networks (Donahue et al., 2018) and manual recording (Lee et al., 2018). For manual recording, crowd-sourcing with the required quality control (speakers who say "cyclohexane" correctly) is prohibitively expensive. As an alternative, we generate a data-set with Google Text-to-Speech on 96,000 factoid questions from a trivia game called Quizbowl (Boyd-Graber et al., 2018), each with 4–6 sentences for a total of over 500,000 sentences.[3] We then decode these utterances using the Kaldi chain model (Peddinti et al., 2015), trained on the Fischer-English dataset (Cieri et al., 2004) for consistency with past results on mitigating ASR errors in MT (Sperber et al., 2017). This model decodes enough noise into our data to test mitigation strategies.[4]

[3] `http://cloud.google.com/text-to-speech`

[4] This model has a Word Error Rate (WER) of 15.60% on the eval2000 test set (Jurafsky et al., 1997). The WER increases to 51.76% on our QB data, which contains out of domain vocabulary. Since there is no past work in question answering, we use machine translation as proxy for determining an appropriate Word Error Rate, as intentional noise has been added to this subdomain (Michel and Neubig, 2018; Belinkov and Bisk, 2018). The most BLEU improvement in machine translation under noisy conditions could be found in this middle WER range, rather than in values below 20% or above 80% (Sperber et al., 2017). Retraining the model on the QB domain would mitigate this

3.2.1 Why Question Answering is challenging for ASR

Question Answering (QA) requires the system to provide a correct answer out of many candidates based on the question's wording. ASR changes the features of the recognized text in several important ways: the overall vocabulary is quite different and important words are corrupted. First, it reduces the overall vocabulary. In our dataset, the number of unique words drops from 263,271 in the original data to a mere 33,333. This is expected, as our ASR system only has 42,000 words in its vocab, so the long tail of the Zipf's curve is lost. Second, unique words—which may be central to answering the question—are lost or misinterpreted; over 100,000 of the words in the original data occur only once. Finally, ASR systems tend to unintentionally delete words, which makes the sentences shorter. In our QB data, the average number of words decreases from 21.62 to 18.85 per sentence.

The decoding system expresses uncertainty by predicting $<unk>$. These account for slightly less than 10% of all our word tokens, but is a top-2 prediction for 30% of the 263,271 words in our dataset. For QA, words with a high TF-IDF measure are valuable. While some words are lost, others can likely be recovered: "hellblazer' becoming "blazer", "clarendon" becoming "claritin". We evaluate this by fitting a TF-IDF model on the Wikipedia dataset and then comparing the average TF-IDF per sentence between the original and the ASR data. The average TF-IDF score drops from 3.52 to 2.77 per sentence, meaning that on average the amount of unique words has decreased. Examples of this change can be seen in Figure 3.1.

noise; however, in practice one is often at the mercy of a pre-trained recognition model due to changes in vocabularies or speakers.

For generalization, we test the effect of noise on two types of distinct questions. QB questions, which are generally four to six sentences long, test a user's depth of knowledge; early clues are challenging and obscure but they progressively become easy and well-known. Competitors can answer these types of questions at any point. Computer QA is competitive with the top players (Yamada et al., 2018). Jeopardy! questions are single sentences and can only be answered after the question ends. To test this alternate syntax, we use the same method of data generation on a dataset of over 200,000 Jeopardy questions (Dunn et al., 2017).

3.3 Mitigating Noise

This section discusses two approaches to mitigating the effects of missing and corrupted information caused by ASR systems. The first approach—forced decoding—exploits systematic errors to arrive at the correct answer. The second uses confidence information from the ASR system to down-weight the influence of low-confidence terms. Both approaches improve accuracy over a baseline DAN model and show promise for short single-sentence questions. However, a third IR approach, specifically using an inverted search index, is more effective on long questions since noisy words are completely avoided during the answer selection process.

3.3.1 IR Baseline

The IR baseline reframes Jeopardy! and QB QA tasks as document retrieval tasks with an inverted search index. We create one document per distinct answer;

each document has a text field formed by concatenating all questions with that answer together. At test time new, unseen questions are treated as queries, and documents are scored using BM25 (Ramos, 2003; Robertson et al., 2009). We implement this baseline with Elastic Search and Apache Lucene.

3.3.2 Forced Decoding

We have systematically lost information due to ASR decoding. We could predict the answer if we had access to certain words in the original question and further postulate that wrong guesses are better than knowing that a word is unknown. For example, "Language is a process of recreation [free creation]" is possible to decipher, while "Language is a process of <*unknown*>" is not.[5]

As a first step, we explored commercial solutions—Bing, Google, IBM, Wit—with low transcription errors. However, their APIs ensure that an end-user often cannot extract anything more than one-best transcriptions, along with an aggregate confidence for the sentence. Additionally, the proprietary systems are moving targets, harming reproducibility.

Therefore, we use Kaldi (Povey et al., 2011) for all experiments. Kaldi is a commonly-used, open-source tool for ASR; its maximal transparency enables approaches that incorporate uncertainty into downstream models. Kaldi provides not only top-1 predictions, but also confidences of words, entire lattices, and phones (Table 3.1). Each item in the sequence represents a word and has corresponding

[5]Providing the full lattice, as in Table 3.1, would grant even more information to the model. However, we did not see an improvement from using the full lattice, likely due to the increased complexity of the data.

Clean	For 10 points, name this revenge novel centering on Edmond Dantes, written by Alexandre Dumas
1-Best	for$^{0.935}$ ten$^{0.935}$ points$^{0.871}$ same$^{0.617}$ this^{1} ...revenge novel centering on <*unk*> written by alexander <*unk*> ...
"Lattice"	for$^{0.935}$ [eps]$^{0.064}$ pretend$^{0.001}$ ten$^{0.935}$ ...pretend point points point name same named name names this revenge novel ...
Phones	f_B$^{0.935}$ er_E$^{0.935}$ t_B$^{0.935}$ eh_I^{1} n_E$^{0.935}$ ...p_B oy_I n_I t_I s_E sil s_B ey_I m_E dh_B ih_I s_E r_B iy_I v_I eh_I n_I jh_E n_B aa_I v_I ah_I l_I ...

Table 3.1: As original data are translated through ASR, it degrades in quality. One-best output captures per-word confidence. Full lattices provide additional words and phone data captures the raw ASR sounds. Our confidence model and forced decoding approach could be used for such data in future work.

confidence in range [0, 1].

The typical end-use of an ASR system wants to know when when a word is not recognized. Under the hood, the ASR system is a graph of possible phrases. In addition tokens for decoded sounds (e.g., "oy", "ah", ...), the graph will have a token that represents an unknown; in Kaldi, this becomes <*unk*>. At a human-level, one would want to know that an out of context word happened.

However, when the end-user is a downstream model, a systematically wrong prediction may be better than a generic statement of uncertainty. So by removing all reference to <*unk*> in the model, we force the system to decode "Louis Vampas" as "Louisiana" rather than <*unk*>.[6] The risk we run with this method is introducing words not present in the original data. For example, "count" and "mount" are similar in sound but not in context embeddings. Hence, we need a method to downweight incorrect decodings.

[6]More specifically, <*unk*> is removed from the Finite State Transducer, which sets the input/output for the ASR system.

3.3.3 Confidence Augmented DAN

We modify the original DAN model (Section 2.3.3) to use word-level confidences from the ASR system as a feature and be robust to corrupted phrases stemming from these incorrect decodings. In increasing order of complexity, the variations are: a Confidence Informed Softmax DAN, a Confidence Weighted Average DAN, and a Word-Level Confidence DAN. We represent the confidences as a vector $\mathbf{c}$, where each cell c_i contains the ASR confidence of word w_i.

The simplest model averages the confidence across the whole sentence and adds it as a feature to the final output classifier. For example in Table 3.1, "for ten points" averages to 0.914. We introduce an additional weight in the output $\mathbf{W^c}$, which adjusts our prediction based on the average confidence of each word in the question. This phrase will not affect the question answering system. But, the following words "revenge novel" have high enough confidences to be decoded, while "Dumas" drops enough to become "$<unk>$".

However, most words have high confidence, and thus the average confidence of a sentence or question level is high. To focus on *which* words are uncertain we weight the word embeddings by their confidence attenuating uncertain words before calculating the DAN average. In the previous example—"for ten points"—"for" and "ten" are frequently occuring words and have a confidence of .935, while "points" has a lower confidence of .871. The next word—"same"—should be "name" and hence the embedding referenced is incorrect. But, the lower confidence of .617 for this prediction decreases the overall weight of the embedding in the model.

Weighting by the confidence directly removes uncertain words, but this is too blunt an instrument, and could end up erasing useful information contained in low-confidence words, so we instead learn a function based on the raw confidence from our ASR system. Thus, we recalibrate the confidence through a learned function f:

$$f(\mathbf{c}) = \mathbf{W^{(c)}c} + \mathbf{b^{(c)}} \tag{3.1}$$

and then use that scalar in the weighted mean of the DAN representation layer:

$$\mathbf{r^{**}} = \frac{\sum_i^N \mathbf{E}[w_{\mathrm{i}}] * f(c_{\mathrm{i}})}{N}. \tag{3.2}$$

In this model, we replace the original encoder $\mathbf{r}$ with the new version $\mathbf{r^{**}}$ to learn a transformation of the ASR confidence that down-weights uncertain words and up-weights certain words. This final model is called our "Confidence Model".

Architectural decisions are determined by hyperparameter sweeps. They include: having a single hidden layer of 1000 dimensionality for the DAN, multiple drop-out, batch-norm layers, and a scheduled ADAM optimizer. Our DAN models train until convergence, as determined by early-stopping. Code is implemented in PyTorch (Paszke et al., 2017), with TorchText for batching.[7]

3.4 Results

Achieving 100% accuracy on this dataset is not a realistic goal, as not all test questions are answerable (specifically, some answers do not occur in the training

[7]Code, data, and additional analysis available at `https://github.com/DenisPeskov/QBASR`

data and hence cannot be learned by an IR-like system). Baselines for the DAN (Table 3.2) establish realistic goals: a DAN trained and evaluated on the *same train and dev set*, only in the original non-ASR form, correctly predicts 54% of the answers. Noise drops this to 44% with the best IR model and down to $\approx$ 30% with neural approaches.

Since the noisy data quality makes full recovery unlikely, we view any improvement over the neural model baselines as recovering valuable information. At the question-level, strong IR outperforms the DAN by around 10%.

There is additional motivation to investigate this task at the sentence-level. Computers can beat humans at the game by knowing certain questions immediately; the first sentence of the QB question serves as a proxy for this threshold. Our proposed combination of forced decoding with a neural model led to the highest test accuracy results and outperforms the IR one at the sentence level.

A strong TF-IDF IR model can top the best neural model at the multi-sentence question level in QB; multiple sentences are important because they progressively become easier to answer in competitions. However, our models improve accuracy on the shorter first-sentence level of the question. This behavior is expected since IR methods are explicitly designed to disregard noise and can pinpoint the handful of unique words in a long paragraph; conversely they are less accurate when they extract words from a single sentence.

	QB				Jeopardy!	
	Synth		Human		Synth	Human
Model	Start	End	Start	End		
Methods Tested on Clean Data						
IR	0.064	0.544	0.400	1.000	0.190	0.050
DAN	0.080	0.540	0.200	1.000	0.236	0.033
Methods Tested on Corrupted Data						
IR base	0.021	0.442	0.180	0.560	0.079	0.050
DAN	0.035	0.335	0.120	0.440	0.097	0.017
FD	0.032	0.354	0.120	0.440	0.102	0.033
Confidence	0.036	0.374	0.120	0.460	0.095	0.033
FD+Conf	0.041	0.371	0.160	0.440	0.109	0.033

Table 3.2: Both forced decoding (FD) and the best confidence model improve accuracy. Jeopardy only has an At-End-of-Sentence metric, as questions are one sentence in length. Combining the two methods leads to a further joint improvement in certain cases. IR and DAN models trained and evaluated on clean data are provided as a reference point for the ASR data.

Speaker	Text
Base	John Deydras, an insane man who claimed to be Edward II, stirred up trouble when he seized this city's Beaumont Palace.
S1	unk an insane man who claimed to be the second unk trouble when he sees unk beaumont → Richard_I_of_England
S2	john dangerous insane man who claims to be the second stirring up trouble when he sees the city's beaumont → London
S3	unk dangerous insane man who claim to be unk second third of trouble when he sees the city's unk palace → Baghdad

Table 3.3: Variation in different speakers causes different transcriptions of a question on Oxford. The omission or corruption of certain named entities leads to different answer predictions, which are indicated with an arrow.

3.4.1 Qualitative Analysis & Human Data

While the synthetic dataset facilitates large-scale machine learning, we ultimately we care about performance on human data. For QB we record questions read by domain experts at a competition. To account for variation in speech, we record five questions across ten different speakers, varying in gender and age; this set of fifty questions is used as the human test data. Table 3.3 provides examples of

variations. For Jeopardy! we manually parsed a complete episode.

The predictions of the regular DAN and the confidence version can differ. As one example, input about The House on Mango Street, which contains words like "novel", "character", and "childhood" alongside a corrupted name of the author, the regular DAN predicts The Prime of Miss Jean Brodie, while our version predicts the correct answer. As another example the model in Table 3.3 predicts "London" if "beaumont" and "john" are preserved, but "Baghdad" if the proper nouns, but not "palace" and "city", are lost.

3.5 Confidence in Data Quality

Confidences are a readily human-interpretable concept that may help build trust in the output of a system. Transparency in the quality of up-stream content can lead to downstream improvements in a plethora of NLP tasks.

Exploring sequence models or alternate data representations may lead to further improvement. Including full lattices may mirror past results for machine translation (Sperber et al., 2017) for the task of question answering. Using unsupervised approaches for ASR (Wessel and Ney, 2004; Lee et al., 2009) and training ASR models for decoding QB or Jeopardy! words are avenues for further exploration.

3.5.1 Can Question Answering Audio be Automated?

Question answering, like many NLP tasks are impaired by noisy inputs. Introducing ASR into a QA pipeline corrupts the data. A neural model that uses the ASR

system's confidence outputs and systematic forced decoding of words rather than unknowns improves QA accuracy on QB and Jeopardy! questions. Our methods are task agnostic and can be applied to other supervised NLP tasks. Larger *human-recorded* question datasets and alternate model approaches would ensure spoken questions are answered accurately, allowing human and computer trivia players to compete on an equal playing field. Text-to-Speech technology can create a large dataset, but the unvarying pronunciation, speed, and voice—every single TTS voice is female—ultimately inhibits this approach from being a gold-standard.

We focus on question answering, but speech data can used for other purposes. As one example, speech-to-speech machine translation will require sizable amounts of training data (Zhang et al., 2004). These techniques can ultimately be used in areas such as call centers (Zweig et al., 2006). But in both translation and in a call center dialog, the quality of the speech is important.

3.6 Implications of Automation

The advantages of this method are cost and scalability, which is demanded by the current paradigm of neural models. This however comes at the expense of quality. A limitation of our past work in **automation** is generalization: text-to-speech centers around a handful of primarily female voices that are consistently decoded, while the voices of real humans are decoded with large variations. Unseen data points are likely to confound a model trained on unnatural data. Furthermore, emotionally realistic speech requires appropriate scope, naturalness, and context (Douglas-Cowie

et al., 2003). Automatic methods, such as text-to-speech, will not be able to address these criteria.

Additionally, automated data creation still depends on having quality source data, that often has to come from expert users. In this project, we record **found** questions that were already written by Quizbowl **experts**. Writing hundreds of thousands of our questions would not have been tractable. Hence, expert design is necessary for automation, as implemented in our other automatically-created dataset, which evaluates co-reference (Chapter 4).

Chapter 4: Automatic Data Generation without a Source[1]

Chapter 3 introduces automation for **generating** data. However, in that project there was a source of **found** data. How can one automate data **generation** without an existing source? This chapter uses an **expert** to design a series of rules to automatically **generate** a dataset. Specifically, we create an evaluation dataset for coreference resolution, a sub-task of dialog (Section 2.1.3). The limitations of the automatic data will motivate using crowd-sourcing in Chapter 5. The merit of the **expert** in the process will motivate using them directly in Chapter 6.

4.1 Evaluating Data

Genuinely varied, realistic data is necessary to create models that are robust to minor variations (Neumann et al., 2019). However, equally robust evaluation methodologies are important in ascertaining the quality of the data (Jones, 1994). Current methods, like IAA, focus on quantitative assessments that may inadvertently assess the **annotation**, but not the **generation**, quality of a dataset. Since most datasets are evaluated on the same types of data—SQuAD test data is comparable

[1]Equal effort between Benno Krojer, Dario Stojanovski, Denis Peskov, and supervised by Alex Fraser. 2020. In International Conference on Computational Linguistics. Peskov is responsible for part of template design, selecting concrete nouns for the templates, paper writing, and the video.

to the training data—the linguistic variation of a dataset is not readily captured by standard quantitative metrics like accuracy or F_1. Furthermore, a model that has memorized several key answers upon which it is then tested is not necessarily *learning*; raw analysis of data overlap confirms this risk (Lewis et al., 2020). Datasets meant to effectively and robustly evaluate trained datasets can determine how much of a problem this poses *ex-post-facto*.

As one solution to this limitation, Checklist (Ribeiro et al., 2020) creates a task-agnostic methodology for testing NLP models. The check is done by replacing words with their synonyms and seeing if task accuracy decreases. We extend this work to a specific task in machine translation. There does not exist a dataset that can serve as a **found** source, unlike our past automation work (Chapter 3). The dataset we create is designed by **experts**: specifically native German and native English speakers, and scaled through **automation**. While a similar dataset of the same size could be created without knowledge of either language, the templates used as test data would prove be nonsensical or unnatural.

4.2 Meaningful Model Evaluation in Machine Translation

Machine translation is a classic and complex NLP task that requires diverse linguistic knowledge and data in multiple languages (Section 2.3). Classic datasets were often gathered through extensive collaboration with **experts**. However, recent ones are often created through **crowd-sourcing** or **automatic** methods. Therefore, this is an area well-suited to our evaluation techniques.

We focus on German-English coreference resolution as a representative task. The seemingly straightforward translation of the English pronoun *it* into German requires knowledge at the syntactic, discourse and world knowledge levels for proper pronoun coreference resolution (CR). A German pronoun can have three genders, determined by its antecedent: masculine (*er*), feminine (*sie*) and neuter (*es*). The nuance of this work requires native knowledge of both English and German.

Accuracy in machine translation is at an all-time high with the rise of neural architectures (Wu et al., 2016) but this metric alone is insufficient for evaluation. Previous work (Hardmeier and Federico, 2010; Miculicich Werlen and Popescu-Belis, 2017; Müller et al., 2018) proposes evaluation methods for specifically pronoun translation. Context-aware neural machine translation (NMT) models are capable of using discourse-level information and are prime candidates for this evaluation. We ask:

> Are transformers (Vaswani et al., 2017) truly *learning* this task, or are they exploiting simple heuristics to make a coreference prediction?

To empirically answer this question, we propose extending a contrastive challenge set for automatic English→German pronoun translation evaluation, ContraPro (Müller et al., 2018) (Section 4.6.1), by making small adversarial changes in the contextual sentences.

Our adversarial attacks—inputs that are almost indistinguishable from natural data and yet classified incorrectly by the network (Madry et al., 2017)—on ContraPro show context-aware Transformer NMT models can easily be misled by simple and unimportant changes to the input. However, interpreting the results

obtained from adversarial attacks can be difficult. In our case, trivial changes in language cause incorrect predictions, but both the changes and the prediction would not be noticed by somebody without a mastery of German. NMT uses brittle heuristics to solve CR if trivial changes in pronouns and nouns fool a coreference corpus like ContraPro. However, this will not identify *which* heuristics these are.

For this reason, we propose a new dataset, created from templates (Section 4.7.1), to systematically evaluate which heuristics are being used in coreferential pronoun translation. Inspired by previous work on CR (Raghunathan et al., 2010; Lee et al., 2011) and language model probing (Ettinger et al., 2016), we create templates tailored to evaluating the specific steps of an idealized CR pipeline. We call this collection ContraCAT, **C**ontrastive **C**oreference **A**nalytical **T**emplates. The construction of templates is controlled, enabling us to easily create large number of coherent test examples and provide unambiguous conclusions about the CR capabilities of NMT. While this methodology depends on automation, a technique shown to be unrealistic for speech in Chapter 3, the templates are written in collaborations between a native German speaker and native English speakers. Since automation is subject to quality control issues, this level of expertise is necessary if the adversarial dataset is to be reflective of actual language used by English and German speakers. The procedure used can be adapted to many language pairs with little effort. We also propose a simple data augmentation approach using fine-tuning. This methodology should not change the way CR is handled by NMT and support the hypothesis that automated data techniques have limited applicability. We release a new dataset, ContraCAT, and the adversarial modifications to ContraPro.

ContraCAT applies only to coreference, but the investigation of heuristics is an important research direction in NLP that can measure the issues noted with **automatic** (Chapter 3) and **crowd-sourced** (Chapter 5) datasets. Heuristics are accurate if there are underlying data limitations; this implies that the training data and the evaluation data resemble one another in superficial ones. Therefore, exposing the brittleness in current datasets motivates the need for higher-quality evaluation data—to observe limitations—and varied training data—to overcome them.

We introduce coreference resolution as a task in Section 4.3, the idealized coreference pipeline in Section 4.3, and the transformer model in Section 4.5. We discuss ContraPro in Section 4.6.1, and explain our proposed templates in Section 4.6.2.

4.3 Why is Coreference Resolution Relevant?

Evaluating discourse phenomena is an important first step in evaluating MT. Apart from document-level coherence and cohesion, anaphoric pronoun translation has proven to be an important testing ground for the ability of context-aware NMT to model discourse. Anaphoric pronoun translation is the focus of several works in context-aware NMT (Bawden et al., 2018; Voita et al., 2018; Stojanovski and Fraser, 2018; Miculicich et al., 2018; Voita et al., 2019; Maruf et al., 2019).

The choice of an evaluation metric for CR is nontrivial. BLEU (Papineni et al., 2002) is the standard metric for machine translation that compares sentence similarity at a word level between two sentences. BLEU-based evaluation is insufficient for measuring improvement in CR (Hardmeier, 2012) without carefully selecting

Start:	The cat and the actor were hungry.
Original Sentence	It (?) was hungrier.
Step 1:	The **cat** and the **actor** were hungry.
Markable Detection	It (?) was hungrier.
Step 2:	The cat and the actor were hungry.
Coreference Resolution	**It** was hungrier.
Step 3:	Der Schauspieler und die Katze waren
Language Translation	hungrig.
	Er / **Sie** / Es war hungriger.

Table 4.1: A hypothetical CR pipeline that sequentially resolves and translates a pronoun.

or modifying test sentences for pronoun translation (Voita et al., 2018; Stojanovski and Fraser, 2018). Alternatives to BLEU include F_1, partial credit, and oracle-guided approaches (Hardmeier and Federico, 2010; Guillou and Hardmeier, 2016; Miculicich Werlen and Popescu-Belis, 2017). However, Guillou and Hardmeier (2018) show that these metrics can miss important cases and propose semi-automatic evaluation. In contrast, our evaluation will be *completely* automatic.

We focus on scoring-based evaluation (Sennrich, 2017), which works by creating contrasting pairs and comparing model scores. Accuracy is calculated as how often the model chooses the correct translation from a pool of alternative minimal edit distance incorrect translations.[2] We are able to scale the size of our adversarial evaluation due to the metric being automatic.

Our work is related to adversarial datasets for testing robustness used in Natural Language Processing tasks such as studying gender bias (Zhao et al., 2018; Rudinger et al., 2018; Stanovsky et al., 2019), natural language inference (Glockner et al., 2018) and classification (Wang et al., 2019b).

[2]Specifically, these alternatives are the two other possible pronouns in German.

4.4 Do Androids Dream of Coreference Translation Pipelines?

Imagine a hypothetical coreference pipeline that generates a pronoun in a target language, as illustrated in Table 4.1. *First*, tag markables, entities that can be referred to by pronouns, in the source sentence since semantics affect binding (Bach and Partee, 1980).[3] Then, detect the subset of animate entities, and separate human entities from other animate ones, since *it* usually cannot refer to a human entit. *Second*, resolve coreferences in the source language. This entails addressing phenomena such as world knowledge, pleonastic *it*, and event references. *Third*, translate the pronoun into the target language. This requires selecting the correct gender given the referent (if there is one), and selecting the correct grammatical case for the target context (e.g., accusative, if the pronoun is the grammatical object in the target language sentence).

This idealized pipeline would produce the correct pronoun in the target language *and* allow a human to understand why the pronoun decision was made. These coreference steps resemble the rule-based approach implemented in Stanford CoreNLP's CorefAnnotator (Raghunathan et al., 2010; Lee et al., 2011) and superficially resemble the three-pronged formulation of Discourse Prominence Theory (Gordon and Hendrick, 1998). However, NMT models are unable to decouple the individual steps of this pipeline, even if they are able to produce the correct pronoun. We propose to isolate each of these steps through targeted examples to understand where the NMT made its decision.

[3]We restrict ourselves to concrete entities as concepts are incompatible with many verbs.

4.5 Model

We use a transformer model (Section 2.3.4) for all experiments. The context-aware model in our experimental setup is a concatenation model (Tiedemann and Scherrer, 2017) (CONCAT) which is trained on a concatenation of consecutive sentences. CONCAT is a standard transformer model and it differs from the sentence-level model only in the way that the training data is supplied to it. Previously, attention-based models discarded information outside of sentence boundaries. Tiedemann and Scherrer (2017) do not modify the model architecture but concatenate preceding and subsequent sentences to the sentence being translated. We train a sentence-level model without any additional concatenation as a baseline.[4]

4.6 ContraPro: Adversarial Attacks on an Adversarial Dataset

ContraPro (Müller et al., 2018), a contrastive challenge set (Section 4.2), has limitations that our new dataset, ContraCAT, will address.

4.6.1 About ContraPro

ContraPro is a contrastive challenge set for English→German pronoun translation evaluation. The set consists of English sentences containing an anaphoric pronoun *it* and the corresponding German translations (e.g., "*Give me your hand,*

[4]The training examples for this model are modified by prepending the previous source and target sentence to the main source and target sentence. The previous sentence is separated from the main sentence with a special token <SEP>, on both the source and target side. This also applies to how we prepare the ContraPro and ContraCAT data. We train the concatenation model on OpenSubtitles2018 data prepared in this way. We remove documents overlapping with ContraPro.

ah, it's soft and hot, and it feels pleasant"→"*Gib deine Hand, ah, sie ist weich und warm, und wohlig fühlt sie sich an.*"). It contains three contrastive translations, differing based on the gender of the translation of *it*: *er*, *sie*, or *es*. The challenge set artificially balances the amount of sentences where *it* is translated to each of these three German pronouns. The appropriate antecedent may be in the main sentence or in a previous sentence. For evaluation, a model needs to produce scores for all three possible translations, which are compared against ContraPro's gold labels.

There may be an inherent skew in the data since it is **found** in movie dialogues rather than being **generated** for specifically testing neural coreference resolution. To ferret out any bias learned by the model from this training set , we generate automatic adversarial attacks on ContraPro that modify the theoretically inconsequential parts of the context sentence before the occurrence of *it*. Coreference accuracy degrades from this adversarial attack suggesting that our transformer model is affected by inconsequential priming and that the original dataset did not have an equal distribution across the three pronouns.

4.6.2 Adversarial Attack Generation

Our three modifications, summarized here, are explained in detail in the following sections:

1. **Phrase Addition**: Appending and prepending phrases containing implausible antecedents:

   The Church is merciful *but that's not the point*. It always welcomes the mis-

guided lamb.

2. **Possessive Extension**: Extending original antecedent with possessive noun phrase:

   I hear ~~her~~ *the doctor's* voice! It resounds to me from heights and chasms a thousand times!

3. **Synonym Replacement**: Replacing original German antecedent with synonym of different gender:

   The curtain rises. It rises. → ~~Der Vorhang~~ *Die Gardine* geht hoch. ~~Er~~ *Sie* geht hoch.[5]

Phrase Addition can be applied to all 12,000 ContraPro examples. The second and third attack can only be applied to 3,838 and 1,531 examples, due to the required sentence contingencies.

4.6.2.1 Phrase Addition

This attack modifies the previous sentence by appending phrases such as "*... but he wasn't sure*" and also prepending phrases such as "*it is true*:...". A range of other simple phrases can be used, which we leave out for simplicity. All phrases we tried provided lower scores. These attacks either introduce a human entity or an event reference *it* (e.g., *"it is true"*) which are both not plausible antecedents for the anaphoric *it*.

[5] *der Vorhang* (masc.) and *die Gardine* (fem.) are synonyms meaning *curtain*

4.6.2.2 Possessive Extension

This attack introduces a new human entity by extending the original antecedent A with a possessive noun phrase e.g., "*the woman's A*". Only two-thirds of the 12,000 ContraPro sentences are linked to an antecedent phrase. Grammar and misannotated antecedents exclude half of the remaining phrases. We put POS-tag constraints on the antecedent phrases before extending them. This filters our subset to 3,838 modified examples. Our possessive extensions can be humans (e.g., *the woman's*), organizations (e.g., *the company's*) and names (e.g., *Maria's*) and each is applied to the pertinent examples.

4.6.2.3 Synonym Replacement

The Synonym Replacement attack gets to the core of whether NMT uses CR heuristics as understanding the pronoun-noun relationship is paramount to predicting the correct pronoun. This attack modifies the original German antecedent by replacing it with a German synonym of a different gender. For this we first identify the English antecedent and its most frequent synset in WordNet (Miller, 1995b). We obtain a German synonym by mapping this WordNet synsets to GermaNet (Hamp and Feldweg, 1997) synsets. Finally, we modify the correct German pronoun translation to correspond to the gender of the antecedent synonym. Approximately one quarter of the nouns in our ContraPro examples are found in GermaNet; in 1,531 of these cases, a synonym of different gender could be identified.

[6]The adversarial attacks modify the context, therefore the baseline model's results on the attacks are unchanged and we omit them. Results for Phrase Addition are computed based on all 12,000

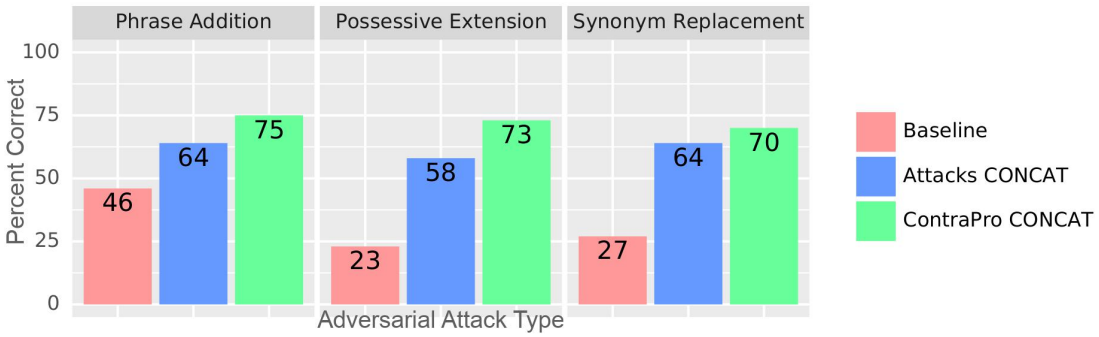

Figure 4.1: The CONCAT model predicts a lower percentage of coreferences correctly when faced with our three adversarial ContraPro attacks. "Attacks CONCAT" shows the drop that our adversarial templates have on "ContraPro CONCAT". **Phrase**: prepending "it is true: ...". **Possessive**: replacing original antecedent *A* with "Maria's *A*". **Synonym**: replacing the original antecedent with different-gender synonyms.[6]

4.6.3 Quality Assessment of the Automatic Attacks by an Expert

We evaluate a random sample of 100 auto-modified examples as a quality control metric. There are 11 issues with semantically-inappropriate synonyms. The model switches from correct to incorrect predictions because of synonym-replacement in 10 of the remaining 89 appropriate examples.[7]

A correct synonym replacement example is:

Es gab einen Brief. Und er war von Sergis Bauer. →

Es gab ein Schreiben. Und es war von Sergis Bauer.

which both mean *"There was a letter. It was from Sergis Bauer."*

One such incorrect synonym replacement that German expert evaluation uncovered is:

ContraPro examples, while for Possessive Extension and Synonym Replacement we only use the suitable subsets of 3,838 and 1,531 ContraPro examples.

[7]Four switches occur from semantically-inappropriate synonyms.

Mein Tisch war so schön gedeckt. Oh, er war hübsch. →

Meine Tabelle war so schön gedeckt. Oh, sie war hübsch.

which means *"My table was neatly decorated. It was pretty".* Both *Tisch* and *Tabelle* translate to *table*, but one is furniture while the other is a matrix. This does not undermine the coreference evaluation itself, since the antecedent is correctly referenced, but it does unintentionally create a semantically implausible sentence.[8]

4.6.4 Evaluating Adversarial Attacks

Intuitively, the adversarial attacks should not contribute to large drops in scores, since no meaningful changes are being made. If the model accuracy drops some, but not all the way to the original sentence-level baseline (Section 4.5), we can conclude that the concatenation model handles CR, but likely with brittle heuristics. If the model accuracy drops all the way to the baseline, then the model is memorizing the inputs. The changes in accuracy suggest issues, but do not ascertain what they are. This problem in pronoun translation evaluation cannot be addressed with simple adversarial attacks on existing general-purpose challenge sets.

4.7 ContraCAT: A Fine-Grained Adversarial Dataset

We propose ContraCAT, a more systematic approach that targets each of the previously outlined CR pipeline steps with data synthetically generated from corresponding templates.

[8] Our templates do not allow for antecedent disagreement to be created in the first place, so there are no direct coreference issues.

Automatic adversarial attacks offer less freedom than templates as many systematic modifications cannot be applied to the average sentence. Thus, our ContraCAT templates are built on the hypothetical coreference pipeline in Section 4.4 that target each of the three steps: 1) Markable Detection, 2) Coreference Resolution and 3) Language Translation. Our minimalistic templates draw entities from sets of animals, human professions (McCoy et al., 2019), foods, and drinks, along with associated verbs and attributes. We use these sets to fill slots in our templates. Animals and foods are natural choices for subject and object slots referenced by *it*. Restricting our sets to interrelated concepts with generically applicable verbs—all animals eat and drink—ensures semantic plausibility. Other object sets, such as buildings, would cause semantic implausibility with certain verbs.

4.7.1 Template Generation

Our templates consist of a previous sentence that introduces at least one entity and a main sentence containing the pronoun *it*. We use contrastive evaluation to judge anaphoric pronoun translation accuracy for each template; we create three translated versions for each German gender corresponding to an English sentence, e.g., *"The cat ate the egg. It rained."* and the corresponding *"Die Katze hat das Ei gegessen.* Er/Sie/Es *regnete"*. To fill a template, we only draw pairs of entities with two different genders, i.e., for animal a and food f: $\text{gender}(a) \neq \text{gender}(f)$. This way we can determine whether the model has picked the right antecedent.

First, we create templates that analyze priors of the model for choosing a

Template Target	Example
Priors	
Grammatical Role	The ***cat*** ate the ***egg***. It (*cat*/*egg*) was big.
Order	I stood in front of the ***cat*** and the ***dog***. It (*cat*/*dog*) was big.
Verb	Wow! She unlocked it.
Markable Detection	
Filter Humans	The ***cat*** and the *actress* were happy. However it (*cat*) was happier.
Coreference Resolution	
Lexical Overlap	The ***cat*** ate the apple and the *owl* drank the water. It (*cat*/ *dog*) ate the apple quickly.
World Knowledge	The ***cat*** ate the *cookie*. It (*cat*) was hungry.
Pleonastic it	The *cat* ate the *sausage*. It was raining.
Event Reference	The *cat* ate the *carrot*. It came as a surprise.
Language Translation	
Antecedent Gender	I saw a ***cat***. It(*cat*) was big. → Ich habe eine Katze gesehen. Sie (*cat*) war groß.

Table 4.2: Template examples targeting different CR steps and substeps. For German, we create three versions with *er*, *sie*, or *es* as different translations of *it*.

pronoun when no correct translation is obvious. Then, we create templates with correct translations, guided by the three broad coreference steps. Table 4.2 provides examples for our templates.

4.7.2 Priors

Our templates that test prior biases do not have a correct answer but reveal the model's biases. We expose three priors with our templates: 1) grammatical roles prior (e.g., subject) 2) position prior (e.g., first antecedent) and 3) a general prior if no antecedent and only a verb is present.

For the first prior, we create a Grammatical Role template where both subject and object are valid antecedents.

For the second prior, we create a Position template where two objects are enumerated as shown in Table 4.2. We create an additional example where the entities order is reversed and test if there are priors for specific nouns or alternatively positions in the sentence.

For the third prior, we create a Verb template, expecting that certain transitive verbs trigger certain object gender choice. We use 100 frequent transitive verbs and create sentences such as the example in Table 4.2.

4.7.3 Markable Detection with a Humanness Filter

Before doing the actual CR, the model needs to identify all possible entities that *it* can refer to. We construct a template that contains a human and animal which are in principle plausible antecedents, if not for the condition that *it* does not refer to people. For instance, the model should always choose *cat* in " *The* actress *and the* cat *are hungry. However* it *is hungrier.*".

4.7.4 Coreference Resolution

Having determined all possible antecedents, the model chooses the correct one, relying on semantics, syntax, and discourse. The pronoun *it* can in principle be used as an *anaphoric* (referring to entities), *event reference* or *pleonastic* pronoun (Loáiciga et al., 2017). For the anaphoric *it*, we identify two major ways of identifying the antecedent: lexical overlap and world knowledge. Our templates for these categories are meant to be simple and solvable.

Overlap: Broadly speaking the subject, verb, or object can overlap from the previous sentence to the main sentence, as well as combinations of them. This gives us five templates: subject-overlap, verb-overlap, object-overlap, subject-verb-overlap and object-verb-overlap.

We always use the same template for the context sentence, e.g., *"The **cat** ate the apple and the **owl** drank the water."*. For the object-verb-overlap we would then create the main sentence *"It ate the apple quickly."* and expect the model to choose *cat* as antecedent. To keep our overlap templates order-agnostic, we vary the order in the previous sentence by also creating *"The **owl** drank the water and the **cat** ate the apple."*

World Knowledge: CR has been traditionally seen as challenging as it requires world knowledge. Our templates test simple forms of world knowledge by using attributes that either apply to animal or food entities, such as *cooked* for food or *hungry* for animals. We then evaluate whether the model chooses e.g., *cat* in *"The **cat** ate the cookie. It was hungry."*

Pleonastic and Event Templates: For the other two ways of using *it*, event reference and pleonastic-it, we again create a default previous sentence (*"The **cat** ate the* apple*.")*. For the main sentence, we used four typical pleonastic and event reference phrases such as *"It is a shame"* and *"It came as a surprise"*. We expect the model to correctly choose the neuter *es* as a translation every time.

4.7.5 Translation to German

After CR, the decoder has to translate from English to German. In our contrastive scoring approach the translation of the English antecedent to German is already given. However the decoder is still required to know the gender of the German noun to select between *er*, *sie* or, *es*. We test this with a list of concrete nouns selected from Brysbaert et al. (2014), which we filter for nouns that occur more than 30 times in the training data. This selects 2051 nouns that are substituted for N in: *"I saw a N. It was {big, small}."*.

4.7.6 Results

The CONCAT model becomes less accurate when actual CR is required. It frequently falls back to choosing the neuter *es* or preferring a position (e.g., first of two entities) for determining the gender. For **Markable Detection** the model always predicts the neuter *es* regardless of the actual genders of the entities.

In the **Overlap** template, the model fails to recognize the overlap and has a general preference for one of the two clauses. In the case of verb-overlap, the model has an accuracy of 64.1% if the verb overlaps with the first clause (*"The cat ate and the dog drank. It ate a lot."*), but a low accuracy of 39.0% when the verb overlaps with the second clause (*"The cat ate and the dog drank. It drank a lot."*.) The overall accuracy for the overlap templates is 47.2%, with little variation across the types of overlap. Adding more overlap, e.g., by overlapping both the verb and object (*"It ate the apple happily"*), yields no improvement. Overall, the model pays little attention

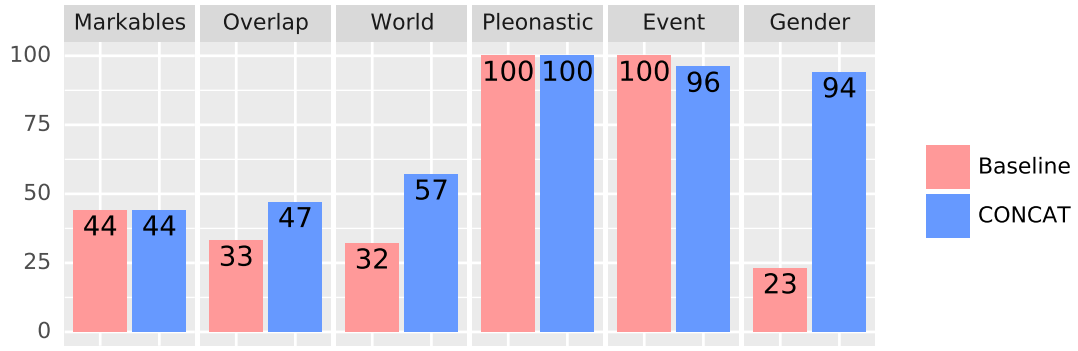

Figure 4.2: Results comparing the sentence-level baseline to CONCAT on ContraCAT. Pronoun translation pertaining to World Knowledge and language-specific Gender Knowledge benefits the most from additional context.

Antecedent-free augmentation	
Source	You let me worry about that. <SEP> How much you take for it?
Reference	Lassen Sie das meine Sorge sein. <SEP> Wie viel kostet er?
Augmentation 1	Lassen Sie das meine Sorge sein. <SEP> Wie viel kostet sie?
Augmentation 2	Lassen Sie das meine Sorge sein. <SEP> Wie viel kostet es?

Table 4.3: Examples of training data augmentations. The source side of the augmented examples remains the same.

to overlaps when resolving pronouns.

The model occasionally predicts answers that require world knowledge, but most predictions are guided by a prior for choosing the neuter *es* or a prior for the subject. An accuracy of 55.7% is slightly above the heuristic of randomly choosing an entity (= 50.0%). This same neuter *es* bias causes the model to have a high accuracy of 96.2% for event reference and pleonastic templates, where *es* is always the correct answer. Based on the high accuracy on the Gender template in Section 4.7.5, we conclude the model consistently memorized the gender of concrete nouns. Hence, CR mistakes stem from Step 1 or Step 2, suggesting that the model failed to learn proper CR.

4.8 Augmentation

We present an approach for augmenting ContraPro to improve CR. Augmentation systematically expands the data to improve a model's robustness (Kafle et al., 2017). While challenging for NLP, we focus on a narrow problem which lends itself to easier data manipulation. Figure 4.2 shows that our model is capable of modeling the gender of nouns. However, there is a strong prior for translating *it* to *es* and hence little intelligent CR capability. Our goal with the augmentation is to alter the prior and test if this can improve CR in the model.

We augment our training data and call it antecedent-free augmentation (AFA). We identify candidates for augmentation as sentences where a coreferential *it* refers to an antecedent not present in the current or previous sentence (e.g., *I told you before. <SEP> It is red. → Ich habe dir schonmal gesagt. <SEP> Es ist rot.*). We create augmentations by adding two new training examples where the gender of the German translation of "it" is modified (e.g., the two new targets are "*Ich habe dir schonmal gesagt. <SEP> Er ist rot.*" and "*Ich habe dir schonmal gesagt. <SEP> Sie ist rot.*"). The source side remains the same. Table 4.3 provides an additional example. Antecedents and coreferential pronouns are identified using a CR tool (Clark and Manning, 2016a,b). We fine-tune our already trained concatenation model on a dataset consisting of the candidates and the augmented samples. As a baseline, we fine-tune on the candidates to confidently say that any potential improvements come from the augmentations.

4.8.1 Augmentation Improves Coreference Accuracy

Augmentation improve coreference accuracy on both ContraPro and ContraCAT. Details are provided in separate sections.

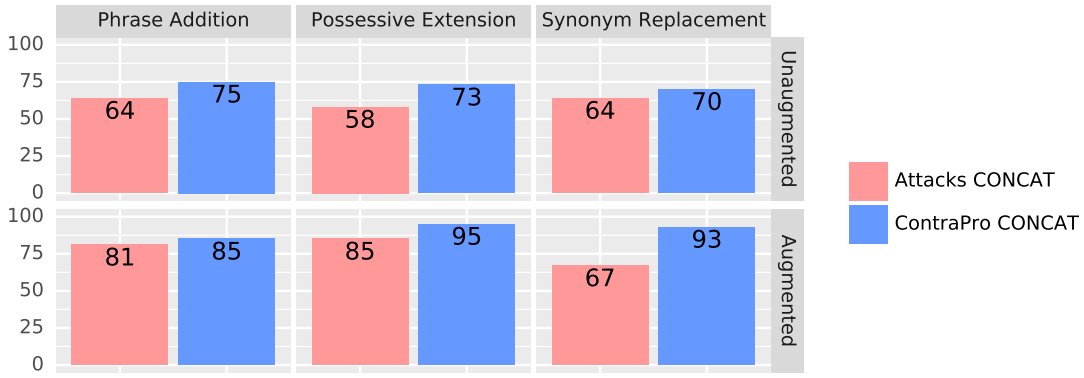

Figure 4.3: Results comparing unaugmented and augmented CONCAT on ContraPro and same 3 attacks as in Figure 4.1. Results with non-augmented CONCAT are the same as Figure 4.1.

4.8.2 ContraPro Results

AFA provides large improvements, scoring 85.3% on ContraPro (Figure 4.3). Since the datasets themselves are slightly different due to the augmentation, we must recompute the baseline. The AFA baseline (fine-tuning on the augmentation candidates only) is higher by 1.94%, presumably because many candidates consist of coreference chains of "it" and the model learns they are important for coreferential pronouns. This improvement in the baseline is small compared to AFA improvements in the full models.

Prediction accuracy on *er* and *sie* is substantially increased, suggesting that the augmentation removes the strong bias towards *es*. Although, the adversarial

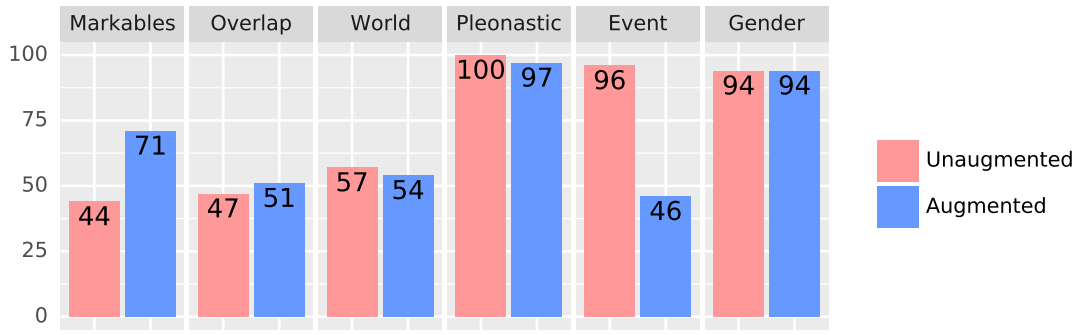

Figure 4.4: ContraCAT results with unaugmented and augmented CONCAT. We speculate that readjusting the prior over genders in augmented CONCAT explains the improvements on Markables and Overlap.

attacks lower AFA scores, in contrast to CONCAT, the model is more robust and the accuracy degradation is substantially lower (except on the synonym attack). We experiment with different learning rates during fine-tuning and present results with the LR that obtain the best baseline ContraPro score. Furthermore, CONCAT and AFA obtain 31.5 and 32.2 BLEU on ContraPro, showing that this fine-tuning procedure, which is tailored to pronoun translation, does not lead to any degradation in overall translation quality.

4.8.3 ContraCAT Results

The prior in ContraCAT over gender pronouns is less concentrated on *es* than in ContraPro. This provides for a more even distribution on the **Position** and **Role Prior** templates.

The augmented model has higher accuracy on **Markable Detection**, improving by 27.6%. Results for the templates are in Figure 4.4.

No improvements are observed on the World Knowledge template. Pleonastic

cases are still accurate, although not perfect as with CONCAT. The Event template identifies a systematic issue with our augmentation. We presume this is due to the CR tool marking cases where *it* refers to events. We do not apply any filtering and augment these cases as well, thus creating wrong examples (an event reference *it* cannot be translated to *er* or *sie*). As a result, the scores are lower compared to CONCAT. This issue with our model is not visible on ContraPro and the adversarial attacks results. In contrast, the Event template easily identifies this problem.

AFA has a similar accuracy to the unaugmented baseline on the Gender template. However, despite increasing by 3.8%, results on Overlap are still underwhelming. Our analysis shows that augmentation helps in changing the prior. We believe this provides for improved CR heuristics which in turn provide for an improvement in coreferential pronoun translation. Nevertheless, the Overlap template shows that augmented models still do not solve CR in a fundamental way.

4.9 Our Dataset in Context

Addressing discourse phenomena is important for high-quality MT. Apart from document-level coherence and cohesion, anaphoric pronoun translation has proven to be an important testing ground for the ability of context-aware NMT to model discourse. Anaphoric pronoun translation is the focus of several works in context-aware NMT (Bawden et al., 2018; Voita et al., 2018; Stojanovski and Fraser, 2018; Miculicich et al., 2018; Voita et al., 2019; Maruf et al., 2019).

Bawden et al. (2018) manually create such a contrastive challenge set for English→French pronoun translation. ContraPro (Müller et al., 2018) follows this work, but creates the challenge set in an automatic way. We show that making small variations in ContraPro substantially changes the accuracy scores, precipitating our new dataset.

Jwalapuram et al. (2019) propose a model for pronoun translation evaluation trained on pairs of sentences consisting of the reference and a system output with differing pronouns. However, as Guillou and Hardmeier (2018) point out, this fails to take into account that often there is not a 1:1 correspondence between pronouns in different languages and that a system translation may be correct despite not containing the exact pronoun in the reference, and incorrect even if containing the pronoun in the reference, because of differences in the translation of the referent. Moreover, introducing a separate model which needs to be trained before evaluation adds an extra layer of complexity in the evaluation setup and makes interpretability more difficult. In contrast, templates can easily be used to pinpoint specific issues of an NMT model. Our templates follow previous work where similar tests are proposed for diagnosing language models (Marelli et al., 2014; Ettinger et al., 2016; Ribeiro et al., 2018; McCoy et al., 2019; Ribeiro et al., 2020).

4.10 Implications for Machine Translation and Automation

In this work, we study how and to what extent CR is handled in context-aware NMT. This work shows that standard challenge sets can easily be manipulated with

adversarial attacks that cause dramatic drops in performance, suggesting that NMT uses a set of heuristics to solve the complex task of CR. Attempting to diagnose the underlying reasons, we propose targeted templates which systematically test the different aspects necessary for CR. This analysis shows that while some type of CR such as pleonastic and event CR are handled well, NMT does not solve the task in an abstract sense. We also propose a data augmentation approach to see if simple data modifications can improve model accuracy. This methodology illustrates the dependence on data by models, and strengthen our claims that low-cost data **generation** techniques are creating datasets that approximate rather than solve NLP tasks. Having identified limitations in existing models, we argue for concrete data extensions for coreference resolution. This methodology—creating an adverserial dataset testing the understanding of a model—can be applied to most NLP tasks.

This project introduces using an **expert**, in this case a native German speaker, in designing the dataset. However, we use templates rather than experts to **automatically** scale the size of the dataset. While we can create *large* datasets, they end up (literally) formulaic. *Solving* tasks like coreference, rather than just noting shortcomings of current datasets, will require building complex and nuanced datasets that allow a model to learn the edge cases of the task. These datasets will ultimately have to built by humans and not **automation**: can the **crowd** be a reliable source of language?

Chapter 5: Crowd-Sourced Generation[1]

Chapters 3 and 4 use **automation** to provide data to solve a task; however, some data cannot be automatically **generated** from templates and require human assistance. **Crowd-sourcing** platforms (Section 2.2.3), specifically Mechanical Turk (Buhrmester et al., 2011), are a cost-efficient, scalable pool for human input. We summarize a data collection project, CANARD, that uses non-**expert** workers to advance question answering (Section 2.1.2) through rewriting trivia questions.

Conversational question answering (CQA) questions differ from machine reading comprehension (MRC) ones in format (Section 2.1.2); however, CQA questions can be rewritten as stand-alone MRC questions to **generate** additional training data. We reduce challenging, interconnected CQA examples to independent, stand-alone MRC to create CANARD—**C**ontext **A**bstraction: **N**ecessary **A**dditional **R**ewritten **D**iscourse—a new dataset that rewrites QUAC (Choi et al., 2018) questions.[2] Language models train on these stand-alone questions with greater flexibility than on CQA ones. Decoupling them allows for new training and test splits. Additionally,

[1] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Peskov is responsible for manual quality control in the data collection process, analysis of the data and model predictions, part of paper writing, and figure&table design.

[2] `http://canard.qanta.org`

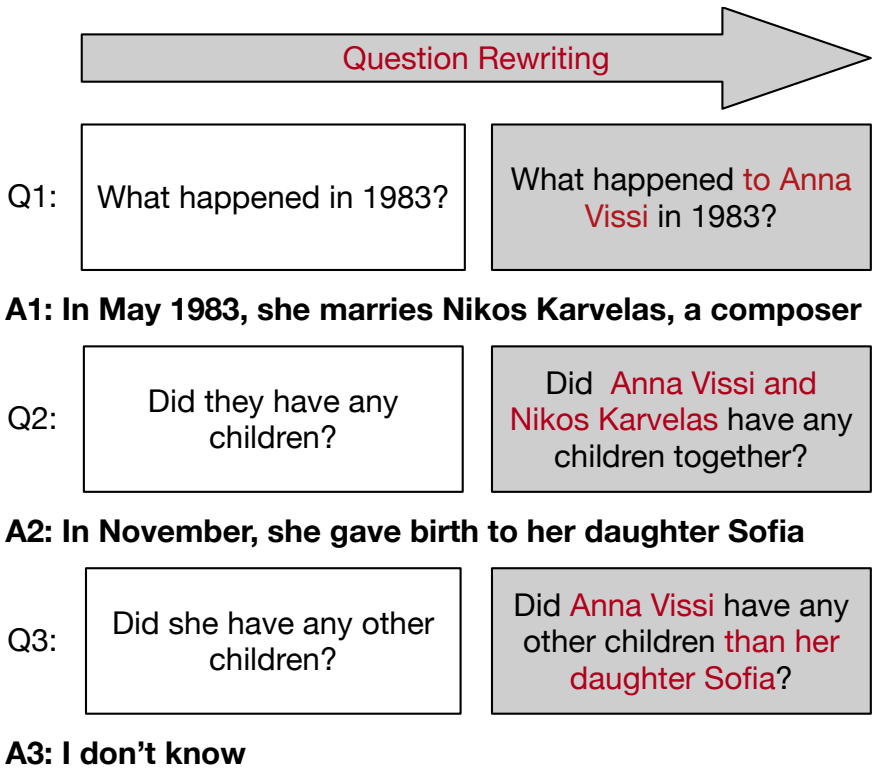

Figure 5.1: Our question-in-context rewriting task. The input to each step is a question to rewrite given the dialog history which consists of the dialog utterances (questions and answers) produced before the given question is asked. The output is an equivalent, context-independent paraphrase of the input question. Crowd-workers are needed to provide these missing details as the omissions are non-formulaic.

successfully rewriting questions to be independent precipitates rewriting questions to be novel. We **crowd-source** context-independent paraphrases of QUAC questions and use the paraphrases to train and evaluate question-in-context rewriting. In the process, we observe the behavior of crowd users and the quality of their output.

Section 5.1 constructs CANARD, a new dataset of question-in-context with corresponding context-independent paraphrases. Section 7.6 analyzes our rewrites (and the underlying methodology) to understand the linguistic phenomena that make CQA and using crowd-sourcing for **generation** difficult.

Characteristic	Ratio
Answer Not Referenced	0.98
Question Meaning Unchanged	0.95
Correct Coreferences	1.0
Grammatical English	1.0
Understandable w/o Context	0.90

Table 5.1: Manual inspection of 50 rewritten context-independent questions from CANARD suggests that the new questions have enough context to be independently understandable.

5.1 Dataset Construction

We elicit paraphrases from human crowdworkers to make previously context-dependent questions *unambiguously* answerable. Through this process, we resolve difficult coreference linkages and create a pair-wise mapping between ambiguous and context-enriched questions. We derive CANARD from QUAC (Choi et al., 2018), a sequential question answering dataset about specific Wikipedia sections. QUAC uses a pair of workers—a "student" and a "teacher"—to ask and respond to questions. The "student" asks questions about a topic based on only the title of the Wikpedia article and the title of the target section. The "teacher" has access to the full Wikipedia section and provides answers by selecting text that answers the question. With this methodology, QUAC gathers 98k questions across 13,594 conversations. We take their entire dev set and a sample of their train set and create a custom JavaScript task in Mechanical Turk that allows workers to rewrite these questions. JavaScript hints help train the users and provide automated, real-time feedback.

We provide workers with a comprehensive set of instructions and task examples (Figure 5.2). We ask them to rewrite the questions in natural sounding English while

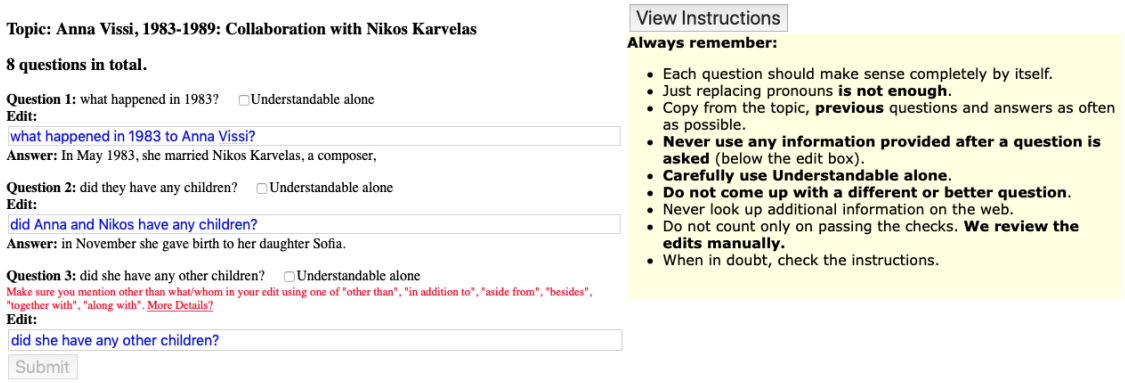

Figure 5.2: The interface for our task guides workers in real-time.

preserving the sentence structure of the original question. We discourage workers from introducing new words that are unmentioned in the previous utterances and ask them to copy phrases when appropriate from the original question. These instructions ensure that the rewrites only resolve conversation-dependent ambiguities. Thus, we encourage workers to create minimal edits.

We display the questions in the conversation one at a time, since the rewrites should include only the previous utterance. After a rewrite to the question is submitted, the answer to the question is displayed. The next question is then displayed. This repeats until the end of the conversation. Figure 5.2 displays the full set of instructions and the data collection interface.

We apply quality control throughout our collection process, given the known **generation** issues (Section 2.2.3). During the task, JavaScript checks automatically monitor and warn about common errors: submissions that are abnormally short (e.g., 'why'), rewrites that still have pronouns (e.g., 'he wrote this album'), or ambiguous words (e.g., 'this article', 'that'). Many QUAC questions ask about 'what/who else' or ask for 'other' or 'another' entity. For that class of questions, we

ask workers to use a phrase such as 'other than', 'in addition to', 'aside from', 'besides', 'together with' or 'along with' with the appropriate context in their rewrite.

We gather and review our data in batches to screen potentially compromised data or low quality workers. A post-processing script flags suspicious rewrites and workers who take and abnormally long or short time. We flag about 15% of our data. *Every* flagged question is manually reviewed by one of the authors and an entire HIT is discarded if one is deemed inadequate. We reject 19.9% of submissions and the rest comprise CANARD. Additionally, we filter out under-performing workers based on these rejections from subsequent batches. To minimize risk, we limit the initial pool of workers to those that have completed 500 HITs with over 90% accuracy and offer competitive payment of $0.50 per HIT.

We verify the efficacy of our quality control through manual review. A random sample of fifty questions sampled from the final dataset is reviewed for desirable characteristics by a native English speaker in Table 5.1. Each of the positive traits occurs in 90% or more of the questions. Based on our sample, our edits retain grammaticality, leave the question meaning unchanged, and use pronouns unambiguously. In one question, a part of the answer is introduced in the rewrite. In five questions, some of the context is under-specified. These infrequent mistakes should not affect our models. We provide examples of failures in Tables 5.2.

We use the rewrites of QUAC's development set as our test set (5,571 question-in-context and corresponding rewrite pairs) and use a 10% sample of QUAC's training set rewrites as our development set (3,418); the rest are training data (31,538).

ORIGINAL: Was this an honest mistake by the media?
REWRITE: Was the claim of media regarding Leblanc's room come to true?
ORIGINAL: What was a single from their album?
REWRITE: What was a single from horslips' *album*?
ORIGINAL: Did they marry?
REWRITE: Did Hannah Arendt and Heidegger marry?

Table 5.2: Not all rewrites correctly encode the context required to answer a question. We take two failures to provide examples of the two common issues: Changed Meaning (top) and Needs Context (middle). We provide an example with no issues (bottom) for comparison.

5.2 Dataset Analysis

We analyze our discuss our datasets with automatic metrics. We compare our dataset to the original QUAC questions and with automatically **generated** questions. We generate the questions with a pronoun substitution baseline that substitutes the Wikipedia title for the pronoun and a simple seq2seq model (Section 2.3.4).[3]

Then, we manually inspect the sources of rewriting errors by the model. Further improvements for the ASR dataset and CANARD are possible.

5.2.1 Anaphora Resolution and Coreference

Our rewrites are longer, contain more nouns and fewer pronouns, and have more word types than the original data. Machine output lies in between the two human-generated corpora, but quality is difficult to assess. Figure 5.3 shows these statistics. We motivate our rewrites by exploring linguistic properties of our data.

[3]We use a bidirectional LSTM encoder-decoder model with shared the word embeddings between the encoder and the decoder. We initialize the embeddings with GloVE (Pennington et al., 2014). We construct the input sequence by concatenating all utterances in the history, prepending them to the message, and adding a special separator token between utterances. Our collected data is split between a training, dev, and test set.

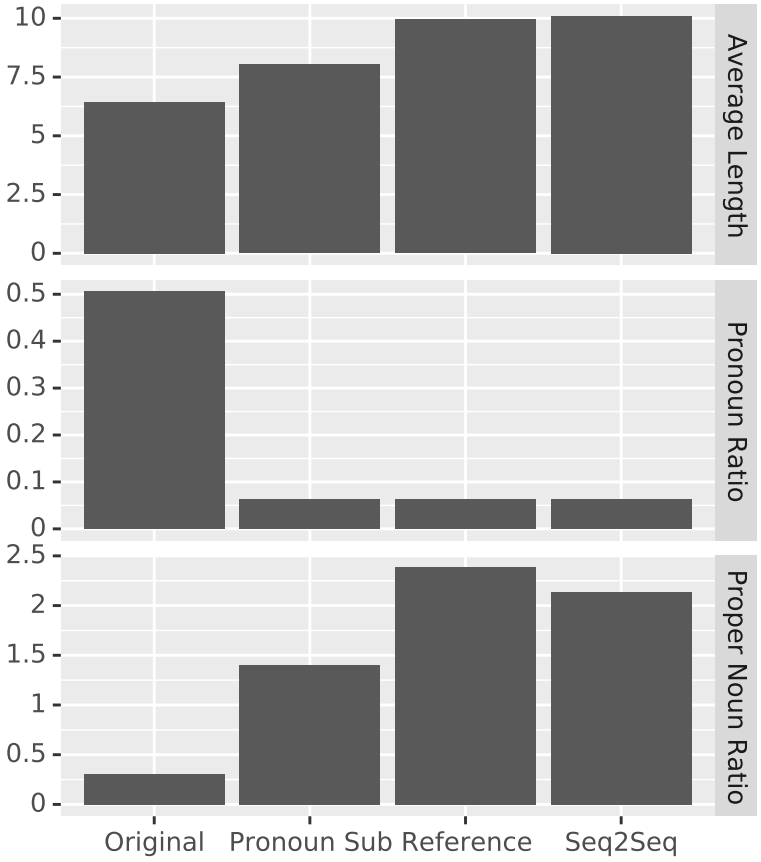

Figure 5.3: Human rewrites are longer, have fewer pronouns, and have more proper nouns than the original QUAC questions. Rewrites are longer and contain more proper nouns than our Pronoun Sub baseline and trained Seq2Seq model.

Anaphora resolution and coreference are two core NLP tasks applicable to this dataset.

Pronouns occur in 53.9% of QUAC questions. Questions with pronouns are more likely to be ambiguous than those without. Only 0.9% of these have pronouns that span more than one category (e.g., 'she' and 'his'). Hence, pronouns within a single sentence are likely unambiguous. However, when viewing the question as an aggregate of sentences, 75.0% of the full questions have pronouns and 27.8% have mixed category pronouns. Therefore, pronoun disambiguation potentially becomes

Label	Text
QUESTION	How long did he stay there?
REWRITE	How long did Cito Gaston stay at the Jays?
HISTORY	*Cito Gaston* **Q:** What did Gaston do after the world series? ... **Q:** Where did he go in 2001? **A:** In 2002, he was hired by the Jays as special assistant to president and chief executive officer Paul Godfrey.

Table 5.3: An example that had over ten flagged proper nouns in the history. Rewriting requires resolving challenging coreferences.

a problem for a quarter of the original data. For example, "Did they argue?" is impossible to answer without context. However, filling this question in with the appropriate context—"Did Johnson and Bird argue?"—allows basketball superfans to answer with a resounding "yes". A full example is provided in Table 5.3.

Approximately one-third of the questions generated by our pronoun-replacement baseline are within 85% string similarity to our rewritten questions. Automatic methods (Chapters 3 and 4) can quickly but somewhat inaccurately replace pronouns with a default phrase. That leaves two-thirds of our data that cannot be solved with pronoun resolution alone. These are unable to be done without a human-in-the-loop.

5.3 Conclusion

Rewriting questions is a challenging stand-alone task and has obvious benefits for question answering. Question rewriting has been formalized as Conversational Question Reformulation (Lin et al., 2020; Vakulenko et al., 2020). (Qu et al., 2020) use CANARD in expanding QuAC for open domain question answering.

More broadly, CANARD is representative of **crowd-sourcing** for **generation**.

The clear limitation of generalist **crowd-sourcing** is the inability to automatically quality control **generated** data. Our work requires *manual* analysis of each sentence submitted by the crowd; this is time-intensive and subject to error. Additionally, it requires real-time task monitoring and user exclusion as otherwise malicious users can quickly contribute a large part of your crowd-sourced task. However, this method generates more diverse and lengthy sentences than comparable **automation** projects (Chapters 3 and 4). One way to handle the quality control issue is by using an **expert** for both **generation** and for quality assessment (Chapter 6).

Chapter 6: Expert Annotation and Evaluation[1]

We introduce a new computational task, adaptation, where the gold standard is subjective and all-important, thereby requiring authoritative **experts**, rather than the anonymous **crowd** (Chapter 5). Vinay and Darbelnet (1995) define adaptation as translation in which the relationship not the literal meaning between the receiver and the content needs to be recreated. The literal translation of named entities such as Anthony Fauci or Dunkin' Donuts into German would keep them the same even if receiver may not have any familiarity with them.

We can use this task to identify named entities (Kasai et al., 2019; Arora et al., 2019; Jain et al., 2019) and for understanding other cultures through creating culturally-centered training data for QA. (Katan and Taibi, 2004). The five-fold task tuple (Section 2.1.1) for adaptation is:

1. **real world problem:** domain adaptation and communication
2. **data:** entities from a given culture
3. **input/output**: an entity from one culture and a corresponding entity belonging to another culture

[1]Denis Peskov, Viktor Hangya, Jordan Boyd-Graber, Alexander Fraser. 2021. In Findings of Empirical Methods in Natural Language Processing, 2021. Peskov is responsible for selecting the entities, designing and running the human generation and the human evaluation, the WikiData work, and writing the paper.

4. **evaluation:** human assessment of relevance in their native culture
5. **standard for progress:** automated adaptations are understandable by relevant humans

We propose two computational methods to find such named entities across American and German culture. However, neither method can be evaluated without a gold standard, which is collected from human annotators. **Annotation** for the human method requires specialized knowledge: familiarity with German or American culture. We use **experts** for this task: domestically educated German and American citizens. Furthermore, **evaluation** requires knowledge of *both* cultures. We hire German translators to assess the computational and human-annotated candidates. This new task is a stepping-stone to automatically generating questions in languages outside of English, the dominant language in the field, and to understanding the perspectives of other cultures (Section 6.6). As one application, health surveys have to be verified for accuracy in translation as medical terminology or demographic may be incorrectly adapted for an audience (Ferrari et al., 2010; Lopes and Trelha, 2013). This chapter explores the use of experts for an **annotation** task. Chapter 7 will use them for **generation**.

6.1 When Translation Misses the Mark

Imagine reading a translation from German, "I saw Merkel eating a Berliner from Dietsch on the ICE". This sentence is opaque without cultural context.

An extreme cultural *adaptation* for an American audience could render the sen-

Bill Gates		
Top Adaptations:		
WikiData	**3CosAdd**	**Human**
F. Zeppelin	congstar	A. Bechtolsheim
Günther Jauch	Alnatura	Dietmar Hopp
N. Harnoncourt	GMX	Carl Benz

Table 6.1: WikiData and unsupervised embeddings (**3CosAdd**) generate adaptations of an entity, such as Bill Gates. Human adaptations are gathered for evaluation. American and German entities are color coded.

tence as "I saw Biden eating a Boston Cream from Dunkin' Donuts on the Acela", elucidating that Merkel is in a similar political post to Biden; that Dietsch (like Dunkin' Donuts) is a mid-range purveyor of baked goods; both Berliners and Boston Creams are filled, sweet pastries named after a city; and ICE and Acela are slightly ritzier high-speed trains.[2] Human translators make this adaptation when it is appropriate to the translation (Gengshen, 2003).

Because adaptation is understudied, we leave the full translation task, which requires **generation**, to future work (Section 6.6). Instead, we focus on the task of cultural adaptation, akin to **annotation**, of entities: given an entity in a source, what is the corresponding entity in English? Most Americans would not recognize Christian Drosten, but the most efficient explanation to an American would be to say that he is the "German Anthony Fauci" (Loh, 2020). We provide top adaptations suggested by algorithms and humans for another American involved with the pandemic response, Bill Gates, in Table 6.1.

Can machines reliably find these analogs with minimal supervision? We generate these adaptations with structured knowledge bases (Section 6.3) and word

[2]We color-code German and American entities throughout.

embeddings (Section 6.4). We elicit human adaptations (Section 6.5.1) to evaluate whether our automatic adaptations are plausible. **Expert** evaluation (Section 6.5.2) validates the merit of our verified annotators relative to computational methods (Section 6.5.3).

6.2 Wer ist Bill Gates?

You could formulate our task as a traditional analogy Drosten::Germany as Fauci::United States (Turney, 2008; Gladkova et al., 2016), but despite this superficial resemblance (explored in Section 6.4), traditional approaches to analogy ignore the influence of culture and are typically *within* a language. Hence, analogies are tightly bound with culture; humans struggle with analogies outside their culture (Freedle, 2003).

Machine translation is another similar task that usually translates words literally; however, this does not necessarily apply in a cultural context as certain named entities may be relevant in one culture but not another. Statistical machine translation (Koehn, 2009) retains an explicit connection between words in the target and source language. In contrast, neural machine translation (Kalchbrenner and Blunsom, 2013) learns a representation of the source in lieu of preserving the original words or phrases. French-English text from the Canadian parliament could be used to train more flexible models than previously possible (Berger et al., 1994). Literature and movie captions (Varga et al., 2007), librettos (Dürr, 2005), medical information (Deléger et al., 2009), and the Internet (Resnik and Smith, 2003; Smith

et al., 2013) can all be sources of parallel data for machine translation.

Creating questions in languages other than English is a current research direction. MLQA and XQuAD automatically generate paired questions through machine translation (Lewis et al., 2019; Artetxe et al., 2019). As an alternative, TyDi (Clark et al., 2020b) gives crowd-sourced users prompts from Wikipedia articles to create questions in a wide range of languages. In all of this work, the goal is to *preserve* the literal meaning of the source as accurately as possible. We propose to *adapt* the meaning to identify new entities and ultimately create new questions.

6.2.1 . . . and why Bill Gates?

This task requires a list of named entities adaptable to other cultures. Our entities come from two sources: a subset of the top 500 most visited German/English Wikipedia pages and the non-official characterization list (Veale, 2016, NOC), "a source of stereotypical knowledge regarding popular culture, famous people (real and fictional) and their trade-mark qualities, behaviours and settings". Wikipedia contains a plethora of singers and actors; we filter the top 500 pages to avoid a pop culture skew.[3] We additionally select all Germans and a subset of Americans from the Veale NOC list as it is human-curated, verified, and contains a broader historical period than popular Wikipedia pages. Like other semantic relationships (Boyd-Graber et al., 2006), this is not symmetric. Thus, we adapt entities in both directions; while Berlin is the German Washington, DC, there is less consensus on what

[3]We discuss the applicability of using Wikipedia (i.e., what proportion of the English Wikipedia is visited from the United States) in Appendix A.1.

is the American Berlin, as Berlin is both the capital, a tech hub, and a film hub. A full list of our entities is provided in Appendix A.2.

6.3 Adaptation from a Knowledge Base

We first adapt entities with a knowledge base. We use WikiData (Vrandečić and Krötzsch, 2014), a structured, human-annotated representation of Wikipedia entities that is actively developed.[4] This resource is well-suited to the task as features are standardized both within and across languages.

Many knowledge bases explicitly encode the nationality of individuals, places, and creative works. Entities in the knowledge base are a discrete sparse vector, where most dimensions are unknown or not applicable (e.g., a building does not have a spouse). For example, Angela Merkel is a human (instance of), German (country of citizenship), politician (occupation), Rotarian (member of), Lutheran (religion), 1.65 meters tall (height), and has a PhD (academic degree). How would we find the "most similar" American adaptation to Angela Merkel? Intuitively, we should find someone whose nationality is American.

Some issues immediately present themselves; contemporary entities will have more non-zero entries than older entities. Some characteristics are more important than others: matching unique attributes like "worked as journalist" is more important than matching "is human".

Each entity in WikiData has "properties", which we can think about as the

[4]We focus on named entities as they are more culturally centered and often have clear location attributes such as place of birth. However, general entities such as "doughnut" and "Berliner" have pages that could be compared.

dimension of a sparse vector and "values" that those properties can take on. For example, Merkel has the properties "occupation" and "academic degree". *Values* for those properties are that her "occupation" is "politician" and her "academic degree" is a "doctorate". To match entities across cultures, we focus on matching properties rather than values; many of the values are more relevant inside a culture. We cannot find American politicians who belong to the Christian Democratic Union, but we can find politicians who have an academic degree and a dissertation title.

As a toy example, if Beethoven, Merkel, and Bach all have only two *properties*: Beethoven has an "occupation" and "genre", Merkel has an "Erdős number" and "political party", and Bach has a "occupation" and "genre", then Beethoven and Bach have a distance of zero from one another and are the closest entities while Merkel has a distance of two since {"Erdős number", "political party"} is two away from {"occupation", "genre"}.

First, we bifurcate WikiData into two sets: an American set $\mathcal{A}$ for items which contain the *value* "United States of America" and a German set $\mathcal{D}$ for those with German values.[5] This is a liberal approximation, but it successfully excludes roughly seven out of the eight million items in WikiData. Then we explore the *properties* from WikiData. We create entity vectors with dimensions corresponding to frequently-occurring properties.

The *properties* are discrete and categorical; Merkel either has an "occupation" or she does not. Each entity then has a sparse vector. We calculate the similarity of

[5]While the geopolitical definition of American is straightforward, the German nation state is more nuanced (Schulze, 1991). Following Green (2003), we adopt members of the Zollverein or the German Confederation as "German" *as well as their predecessor and successor states.* This approach is a more inclusive (Großdeutschland) definition of"German" culture.

the vectors with Faiss's L_2 distance (Johnson et al., 2017) and for each vector in $\mathcal{A}$ find the closest vector in $\mathcal{D}$ and *vice versa*.

So who is the American Angela Merkel? One possible answer is Woodrow Wilson, a member of a "political party", who had a "doctoral advisor" and a "religion", and ended up with "awards". This answer may be unsatisfying as it was Barack Obama who sat across from Merkel for nearly a decade. To capture these more nuanced similarities, we turn to large text corpora in Section 6.4.

6.4 An Alternate Embedding Approach

While the classic NLP vector example (Mikolov et al., 2013d) isn't as magical as initially claimed (Rogers et al., 2017), it provides useful intuition. We can use the intuitions of the cliché:

$$\overrightarrow{\text{King}} - \overrightarrow{\text{Man}} + \overrightarrow{\text{Woman}} = \overrightarrow{\text{Queen}} \tag{6.1}$$

to adapt between languages.

This, however, requires relevant embeddings. First, we use the entire Wikipedia in English and German, preprocessed using Moses (Koehn et al., 2007). We follow Mikolov et al. (2013c) and use named entity recognition (Honnibal et al., 2020) to tokenize entities such as Barack_Obama.

We use word2vec (Mikolov et al., 2013c), rather than FastText (Bojanowski et al., 2017), as we do not want orthography to influence the similarity of entities. Angela Merkel in English and in German have quite different neighbors, and we

intend to keep it that way by preserving the distinction between languages.

However, the standard word2vec model assumes a single monolingual embedding space. We use unsupervised Vecmap (Artetxe et al., 2018), a leading tool for creating cross-lingual word embeddings, to build bilingual word embeddings. We propose two approaches for adaptation.

3CosAdd We follow the word analogy approach of 3CosAdd (Levy and Goldberg, 2014; Köper et al., 2016).[6] American→German adaptation takes the source entity's (v) embedding in the English vector space and looks for its adaptation (u^*) based on embeddings in the German space. This is like the word analogy task, i.e., what entity has the role in the German culture as v does in American culture. As an example, Merkel has a similar role in the German culture as Biden. Formally, the adaptation of the English entity v into German is

$$\overrightarrow{a} \equiv \text{avg}\left(\overrightarrow{E^{en}}_{\text{United\_States}}, \overrightarrow{E^{de}}_{\text{USA}}\right) \tag{6.2}$$

$$\overrightarrow{d} \equiv \text{avg}\left(\overrightarrow{E^{en}}_{\text{Germany}}, \overrightarrow{E^{de}}_{\text{Deutschland}}\right) \tag{6.3}$$

$$u^* = \arg\max_{u \in V^{de}} \text{sim}\left(\overrightarrow{E^{de}_u}, \overrightarrow{E^{en}_v} - \overrightarrow{a} + \overrightarrow{d}\right), \tag{6.4}$$

where $\overrightarrow{E^l_w}$ is the embedding of word w in language l, V^{de} is the German vocabulary and sim is the cosine similarity. The American anchor word $\overrightarrow{a}$ and German anchor $\overrightarrow{d}$ represent the American and German cultures.[7] We average the English and German embeddings of the individual word types for robust anchor vectors.

[6] We experiment with 3CosMul as well but found 3CosAdd generally more robust.

[7] Der Spiegel, the largest newspaper, and other prominent media sources call their United States sections USA.

In standard analogies, as in Equation 6.1, the $\overrightarrow{a}$ and $\overrightarrow{d}$ vectors are different for each test pair; here they are the same for each example, as we always are pivoting between the two cultures.

Learned adaptation To eliminate the need for manual anchor selection for both cultures, our second approach learns the adaptation as a linear transformation of source embeddings to the target culture given a few adaptation examples. Specifically, we use the human adaptations sourced for the Wikipedia entities as training for the Veale NOC ones. We follow the work of Mikolov et al. (2013a) and learn a transformation matrix $\mathbf{W}_{en \to de}$ for American→German by minimizing the L_2 distance of $\mathbf{W}_{en \to de} \overrightarrow{E}^{en}_{v_i}$ and $\overrightarrow{E}^{de}_{u_i}$ over gold adaptation $v_i, u_i{}_{i=1}^{n}$ entity pairs. The adaptation of a source entity v is $u^* = \mathbf{W}_{en \to de} \overrightarrow{E}^{en}_{v}$. Likewise, we learn the reverse mapping $\mathbf{W}_{de \to en}$ for German→American adaptation. This requires supervised training data —but not much (Conneau et al., 2017)—since there are no existing gold labels for these adaptations to serve as an oracle. We collect this data from appropriately qualified **experts** in Section 6.5.

6.5 Comparing Automation to Human Judgment

The computational methods can **generate** entities at scale, but humans have to **evaluate** their relevance.

6.5.1 Adaptation by Locals

Since quality control is difficult for **generation** and complicated **annotation** (Karpinska et al., 2021), we need users who will answer the task accurately. We recruit five American citizens educated at American universities and five German citizens educated at German ones through personal educational networks. They have familiarity with the popular named entities in their own culture, which is the necessary expertise for adaptation. These human **annotations** serve as a gold standard against which we can compare our **automatic** approaches. Chapter 5 showed that human output was superior to the automatic approaches for the notably more straightforward task of question rewriting. To improve the user experience, we create an interface (Figure 6.2) that provides a brief summary of each source entity from Wikipedia and asks the users to select a target adaptation that autocompletes Wikipedia page titles (all entities; targets are not limited to the lists in Section 8.4.2) in a text box *a la* answer selection in Wallace et al. (2019b). We provide a thought-through example of possible adaptations for Angela Merkel in our instructions and encourage a holistic approach to the task. The annotation task requires two hours for our users to complete. Each annotation is independent allowing users to return to the task at their convenience over a span of two weeks. Participants completed the task on a volunteer basis. Obviously, German annotators are more familiar with German culture than the Americans, and vice-versa. Annotators translate into their native language. Since we are focusing on popular entities, they are often known despite the cultural divide, but the introductory paragraph from Wikipedia reminds

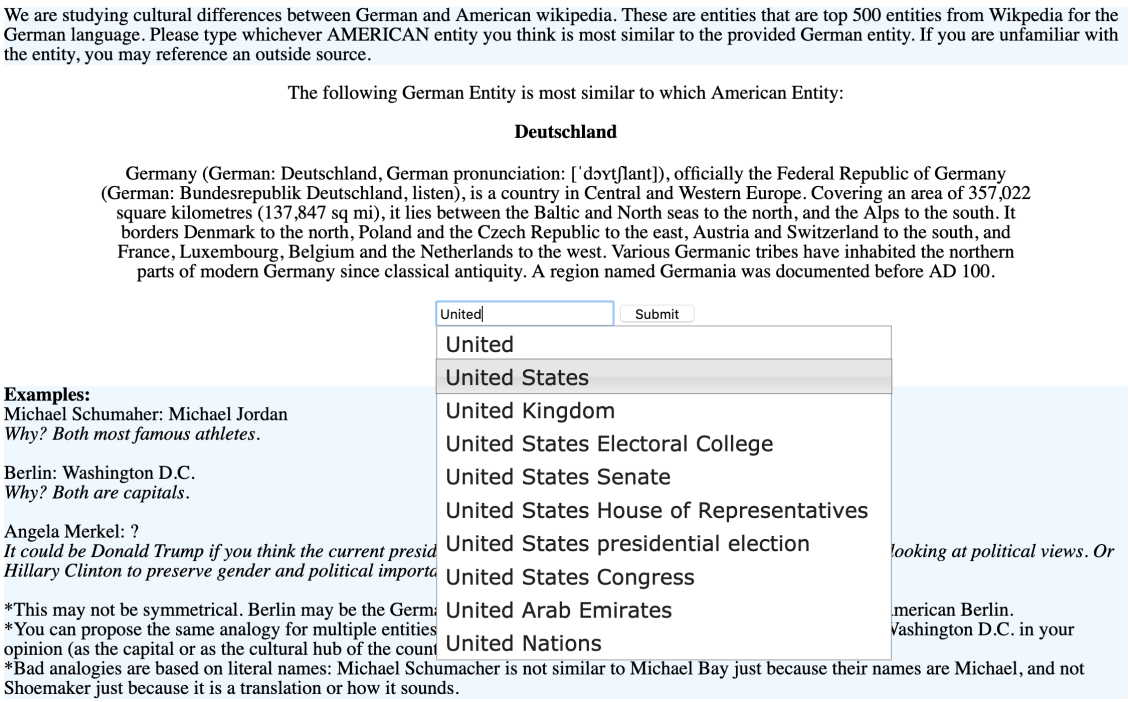

Figure 6.1: Our interface provides users with information about the entity and asks them to select an option from possible Wikipedia pages

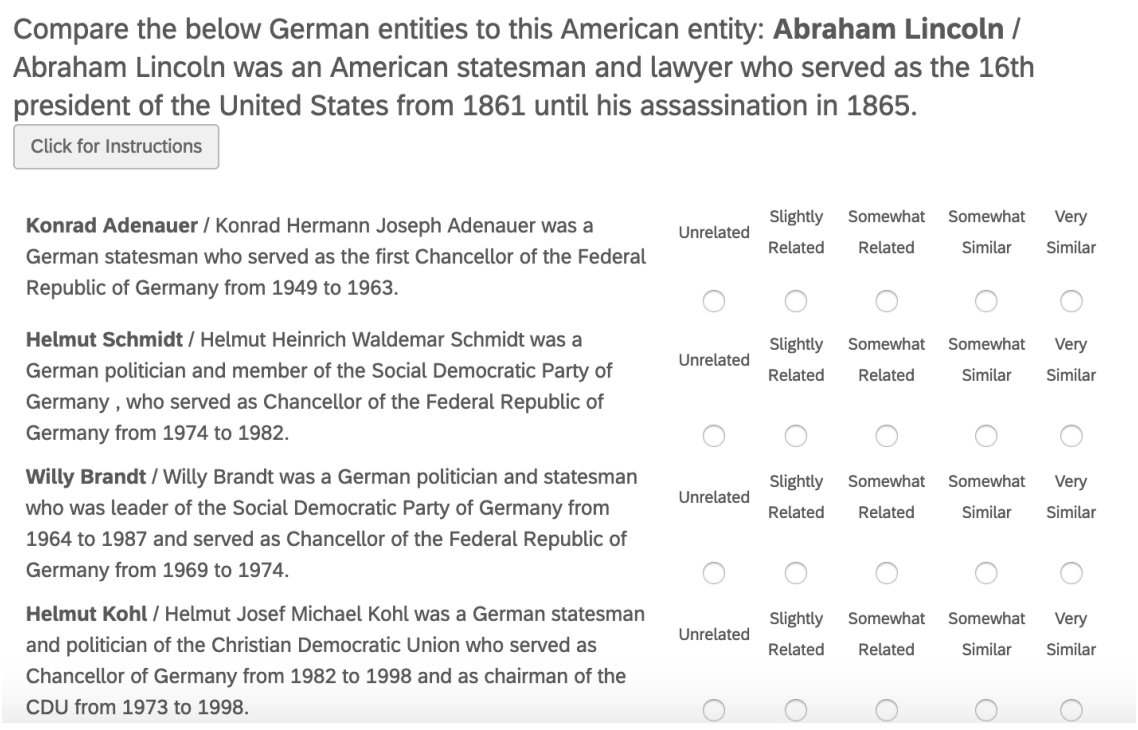

Figure 6.2: Our Qualtrics survey

users if not.

6.5.2 Are the Adaptations Plausible?

To validate and compare all our adaptation strategies' precision, five German translators, appropriately qualified **experts** for the evaluation, who understand American culture assess the adaptations.[8] The top five adaptations from WikiData, 3CosAdd, learned adaptation, and humans—as well as five randomly selected options from the human pool—are evaluated for plausibility on a five-level Likert scale.[9] We provide instructions and examples for using the Likert scale and provide users with a free-response box to escalate concerns. Fleiss' Kappa (0.382) assesses interannotator Agreement; this "fair" agreement suggests that vetting an adaptation is challenging and sometimes subjective, even for translators.

6.5.3 Why Adaptation is Difficult

Embedding adaptations are better than WikiData's, and human adaptations are better still (Figure 6.3). Thus, we use human adaptations as the gold standard for evaluating recall. Only the learned embedding method uses training data, so we use human adaptations from Wikipedia to train the projection matrix and evaluate (for all methods) using human adaptations the NOC list.

Given that the task is subjective, we take our results with a grain of salt given cultural variation (e.g., some people view Angela Merkel's conservatism as a defining characteristic, while others focus on her science pedigree). The adaptations come

[8] Recruited through Upwork for $40 each.

[9] Our custom Qualtrics survey is provided in Figure 6.2. The order of adaptations is randomized and assessed on a Likert scale with anchors from Jurgens et al. (2014).

from verified citizens of the respective countries, which is the appropriate level of **expertise** for this task. Anonymous **crowd** annotation would create unexpected familiarity biases: *all* politicians could be reduced to Angela Merkel and all companies could be reduced to Mercedes-Benz, since there is no obvious mechanism to encourage *great* rather than a *good* annotations.

We use the mean reciprocal rank (Voorhees et al., 1999, MRR) to measure how high the gold adaptations are ranked by our other adaptation strategies. Since MRR decreases geometrically and our gold standard is not exhaustive, the Recall@5, and @100 metrics are more intuitive. We calculate Recall@n by measuring what fraction of the correct adaptations of a source entity is retrieved in the top n predictions.[10] Table 6.2 validates that the human annotations are near the top of the automatic adaptations; the precision-oriented evaluation (Figure 6.3) validates whether the top of the list is reasonable. All human annotations and a sample of the automatic adaptations are provided in Appendix A.2.

6.5.4 Qualitative Analysis

There is no single answer to what makes a good adaptation. Let us return to the question of who Bill Gates is, which underlines how there is often no one right answer to this question but several context-specific possibilities. The human adaptations show the range of plausible adaptations, each appropriate for a particular facet of the position Bill Gates has in US society. As previously mentioned, Carl Benz represents a larger than life founder who created an entire industry with his

[10] This is often referred to as P@n in bilingual lexicon induction literature (Conneau et al., 2017).

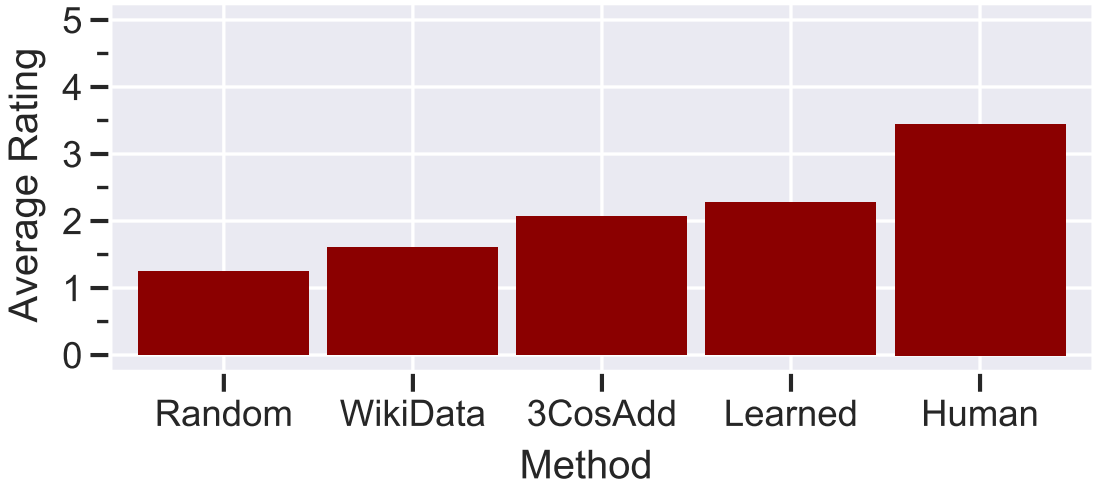

Figure 6.3: We validate adaptation strategies with expert translators on a five-point Likert scale. The human-generated adaptations are rated best—between "related" (3) and "similar" (4). These human adaptations become the reference for evaluation in Table 6.2.

Data	Metric	WikiData	3CosAdd	Learned
American→ *German*				
Wikipedia	Rec@5	7.5%	**14.2%**	-
	Rec@100	34.4%	**52.8%**	-
	MRR	0.05	**0.10**	-
Veale NOC	Rec@5	3.0%	22.9%	**28.6%**
	Rec@100	42.4%	**51.4%**	45.7%
	MRR	0.03	0.17	**0.24**
German→ *American*				
Wikipedia	Rec@5	3.1%	**17.2%**	-
	Rec@100	15.4%	**40.5%**	-
	MRR	0.01	**0.12**	-
Veale NOC	Rec@5	0.0%	**25.0%**	**25.0%**
	Rec@100	25.0%	**70.0%**	55.0%
	MRR	0.02	0.12	**0.15**

Table 6.2: If we consider human adaptations as correct, where do they land in the ranking of automatic adaptation candidates? In this recall-oriented approach, learned mappings (which use a small number of training pairs), rate highest.

company. However, Carl Benz made cars, not computers.

Even within technology, different adaptations highlight different aspects of Bill Gates. Like the implementer of the BASIC programming language, Konrad Zuse

contributed to computers that were more than single-purpose machines. Just as as Bill Gates's Microsoft is seen as a stodgy tech giant, Dietmar Hopp founded SAS, a giant German tech company that is more often discussed in board rooms than in living rooms. And because the epicenter of modern tech is America's West Coast, Andreas von Bechtolsheim represents a German founder of Sun Microsystems and early Google investor that made his way to Silicon Valley.

Other times, there is more consensus: a majority of raters declare Angela Merkel is the German Hilary Clinton, and Joseph Smith is the American Martin Luther. There are even some unanimous adaptations: Bavaria is the German California. Adaptations of fictional characters seem particularly difficult, although this may represent the supremacy of American popular culture; Superman and Homer Simpson are so well known in Germany that there are no clear adaptations; Till Eulenspiegel, Maverick, Bibi Blocksberg are not superheroes from a dying world and Heidi is not a dumb, bald everyman.

We evaluate the translator evaluations as well. The assessment is committed in good faith. Karl Denke, a serial killer and a random control for Abraham Lincoln is rating as "unrelated' (1) by all annotators. The translators generally agree in the direction of the rating even if the exact rating varies: Bismarck for Abraham Lincoln is correctly rated as either a four *or* a five by all annotators; both are historically prominent 19th century politicians responsible for a military unification. However, there are certain differences, such as Abraham Lincoln being heavily associated with his assassination. The overall average is brought down by adaptations such as Napolean for Abraham Lincoln being evaluated as "unrelated' (1) due to not being

Source:	What is the longest river in the United States? Mississippi
Detection:	What is the longest river in the **United States**? **Mississippi**
Adaptation:	What is the longest river in the **Germany**? **Rhine**
Target:	**Welches ist der längste Fluss in Deutschland? Rhein**

Table 6.3: A hypothetical QA pipeline that adapts a question.

a German adaptation, even if the adaptation is otherwise reasonable, which makes a particularly large difference for the human annotations.

6.6 Generating New Questions

These results ultimately bring us back to the motivation: can our methodology be used to generate questions in a new language? We discuss a hypothetical pipeline for doing this and provide an example.

This will require a combination of machine translation and adaptation. First, relevant Named Entities in a sentence must be identified with a Named Entity recognition tool, such as spaCy Part of Speech tagging (Honnibal and Johnson, 2015). Second, these Named Entities must be translated into the target culture. This poses a research challenge, since multiple Named Entities must be translated jointly. Last, the entire sentence must be translated fluently into the target language. This pipeline is illustrated with an example in Table 6.3.

We do not expect that most of our generated questions will make sense; turning "When did Abraham Lincoln make his Emancipation Proclamation" into "When did

Friedrich Ebert make the Edict of Potsdam" is nonsensical.[11] One solution is to decouple the questions and answers. This practice has been recently implemented by Clark et al. (2020b), in which they ask one batch of participants what they are curious about, without any grounding motivation. Then they collect the answers, if they are available, from Wikipedia. This type of approach would allow either half of question answering to be independently adapted. A more complicated approach would require a joint model that is confident in having identified all entities in a sentence, and in coherently adapting them together.

6.6.1 Adaptation is not Trivial

When creating new questions, correct adaptation must navigate complicated political and ideological barriers. Comparing one to Napolean may have completely different connotations in France and in Italy and could cause a political snafu. An incorrect religious comparison could have even higher stakes. Additionally, adaptation may introduce new ideas; in Alice in Wonderland, a character develops the characteristic of being sleepy when the name is adapted into Portuguese (Carroll and Amorim, 2003). Certain adaptations have been made nefariously in the interest of censorship (Tymoczko, 2006). This makes adaptation a useful tool for exploring the abstract idea of culture.

[11] Ebert lived in the twentieth century and could not have authored a seventeenth century edict on religious liberty.

6.7 A New Computational Task

We formally introduce entity **adaptation** as a new computational task and show why **experts** are needed for any subjective task. Word2vec embeddings and WikiData can be used to figuratively—not just literally—translate entities into a different culture. Humans are better at generating candidates for this task than our computational methods (Figure 6.3). These methods are well-motivated, but have room for improvement. Knowledge bases improve over time and increased coverage of entities—as well as improved information about each entity—would improve the method. Alternate word embedding approaches—perhaps those that discard orthography—may provide better candidates. Even humans occasionally disagree with other humans on this task, so evaluation for this task is nontrivial. Since entities have multiple valid adaptations, one cannot exclude adaptations as invalid due to being different from those proposed by other annotators. Hence, excluding an improperly-motivated or improperly qualified *annotator* is more important than excluding *annotations* after the fact.

People need questions and answers that reflect their language **and** culture, but datasets are lacking: adaptation can help. There has been an explosion of English-language QA datasets, but other languages continue to lag behind. Several approaches try to transfer English's bounty to other languages (Lewis et al., 2019; Artetxe et al., 2019), but most of the entities asked about in major QA datasets are American (Gor et al., 2021b). Adapting entire questions will require not just adapting entities and non-entities in tandem but will also require integration with

machine translation (Kim et al., 2019; Hangya and Fraser, 2019). High quality adaptation is paramount to make the questions interesting if they occur in a trivia context and pertinent if they occur in an educational context.

Human input, either as generators or evaluators of the adaptations, is required at this stage for adaptations to be reliable. Our **automatic** methods did not create precise adaptations, but the alternative "incorrect" adaptations may be useful for low-precision tasks, such as generating numerous simple open-ended questions or gauging the popularity of an entity. Additionally, our new dataset of human adaptations and human evaluation of these adaptations can serve as an evaluation metric for future automatic methods. Given the existence of robust datasets in high resource languages can we **adapt**, rather than literally translate, them to other cultures and languages?

This task is not possible without **expert annotation**. However, we do not **generate** full translations in this task. We do not observe malicious or careless answers from our **annotators** or **evaluators**. Hence, we extend the use of **experts** to a task in which quality assurance is nearly impossible: dialog **generation** in Chapter 7.

Chapter 7: Expert Generation[1]

Experts can **generate** datasets of a quality unachievable by the **crowd** by providing *reliable* and *specialized* expertise. First, working with experts usually involves verifying their identify and creating an ongoing relationship, often in the form of a contract. This relationship enables tasks requiring a long-term commitment with a user; a psuedo-anonymous crowd user being paid to complete an independent task does not have strong motivation to consistently repeat the one-time task. Second, specialized knowledge may be needed for certain tasks; a larger amount of incorrect x-ray annotations would not be preferable to a smaller amount of correct ones for a radiologist. Additionally, the accuracy, rather than the size, of the data allows the dataset to withstand the test of time.[2] This justifies the large investment of time, relationship-building, and money necessary to use experts.

We create a deception dataset using experts, as a contrast to the earlier crowdsourced generated CANARD dataset (Chapter 5). Participants—that are engaged in the task and are appropriately compensated—both generate and annotate data in

[1]Denis Peskov, Benny Chang, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It Takes Two to Lie: One to Lie and One to Listen. In Proceedings of The Association for Computational Linguistics. Peskov is responsible for designing the task, gathering the participants, running the games, building half the models, part of the data analysis, the visualizations, and the paper writing.

[2]The Penn Treebank (Marcus et al., 1993), which used graduate students in linguistics and spanned three years in the early 1990s remained influential for years and is referenced in Computational Linguistic courses today.

the span of a game that usually lasts over a month. The **annotation** is more complicated than in our adaptation dataset (Chapter 6) due to being real-time and user-specific. The resulting product is a gold standard of conversational NLP data in quality of language, diversity, and naturalness.

The conversations and annotations thereof would not be possible without experts familiar with the game. Deception is an art, rather than a science (Bavelas et al., 1990; Bell and DePaulo, 1996) and like adaptation (Chapter 6), a subjective task. We recruit top players from the competitive Diplomacy community (Hill, 2014; Chiodini, 2020) and compensate them appropriately for their effort.

7.1 Where Does One Find Long-Term Deception?

A functioning society is impossible without trust. In online text interactions, users are typically trusting (Shneiderman, 2000), but this trust can be betrayed through false identities on dating sites (Toma and Hancock, 2012), spearphishing attacks (Dhamija et al., 2006), sockpuppetry (Kumar et al., 2017) and, more broadly, disinformation campaigns (Kumar and Shah, 2018). Beyond such one-off antisocial acts directed at strangers, deception can also occur in sustained relationships, where it can be strategically combined with truthfulness to advance a long-term objective (Cornwell and Lundgren, 2001; Kaplar and Gordon, 2004).

We introduce a dataset to study the strategic use of deception in long-lasting relationships. We define the task (Section 2.1.1) of deception detection as:

1. **real world problem:** deception (and an extension of dialog 2.1.3)

Message	**Sender's intention**	**Receiver's percep.**
If I were lying to you, I'd smile and say "that sounds great." I'm honest with you because I sincerely thought of us as partners.	Lie	Truth
You agreed to warn me of unexpected moves, then didn't . . . You've revealed things to England without my permission, and then made up a story about it after the fact!	Truth	Truth
. . . I have a reputation in this hobby for being sincere. Not being duplicitous. It has always served me well. . . . If you don't want to work with me, then I can understand that . . .	Lie	Truth
(Germany attacks Italy)		
Well this game just got less fun	Truth	Truth
For you, maybe	Truth	Truth

Table 7.1: An annotated conversation between Italy (white) and Germany (gray) at a moment when their relationship breaks down. Each message is annotated by the sender (and receiver) with its intended or perceived truthfulness; Italy is lying about . . . lying.

2. **data:** raw text that contains elements of deception
3. **input/output**: input of free form text and output of a binary decision on deceptiveness thereof
4. **evaluation:** accuracy of the above decision, relative to humans
5. **standard for progress:** reaching parity with human detection of deception

To collect reliable ground truth in this complex scenario, we design an interface for players to naturally **generate** and **annotate** conversational data while playing a negotiation-based game called Diplomacy. These annotations are done in *real-time* as the players send and receive messages. While this game setup might not directly translate to real-world situations, it enables computational frameworks for studying deception in a complex social context while avoiding privacy issues.

After providing background on the game of Diplomacy and our intended de-

ception annotations (Section 7.2), we discuss our study (Section 7.4). To probe the value of the resulting dataset, we develop lie prediction models (Section 7.5) and analyze their results (Section 7.6). The role of the **expert** is paramount (Section 2.2.4).

7.2 Diplomacy

The Diplomacy board game places a player in the role of one of seven European powers on the eve of World War I. The goal is to conquer a simplified map of Europe by ordering armies in the field against rivals. Victory points determine the success of a player and allow them to build additional armies; the player who can gain and maintain the highest number of points wins.[3] The mechanics of the game are simple and deterministic: armies, represented as figures on a given territory, can only move to adjacent spots and the side with the most armies always wins in a disputed move. The game movements become publicly available to all players after the end of a turn.

Because the game is deterministic and everyone begins with an equal amount of armies, a player cannot win the game without forming alliances with other players—hence the name of the game: Diplomacy. Conquering neighboring territories depends on support from another player's armies. After an alliance has outlived its usefulness, a player often dramatically breaks it to take advantage of their erstwhile ally's vulnerability. Table 7.1 shows the end of one such relationship. As in real

[3]In the parlance of Diplomacy games, points are "supply centers" in specific territories (e.g., London). Having more supply centers allows a player to build more armies and win the game by capturing more than half of the 34 supply centers on the board.

life, to succeed a betrayal must be a surprise to the victim. Thus, players pride themselves on being able to lie *and* detect lies. Our study uses their skill and passion to build a dataset of deception created by battle-hardened diplomats. Senders annotate whether each message they write is an ACTUAL LIE and recipients annotate whether each message received is a SUSPECTED LIE. Further details on the annotation process are in Section 7.4.1.

7.2.1 A game walk-through

Figure 7.1 shows the raw counts of one game in our dataset. But numbers do not tell the whole story. We analyze this case study using rhetorical tactics (Cialdini and Goldstein, 2004), which Oliveira et al. (2017) use to dissect spear phishing e-mails and Anand et al. (2011) apply to persuasive blogs. Mentions of tactics are in italic (e.g., *authority*). For the rest of the paper, we will refer to players via the name of their assigned country.

Through two lie-intense strategies—convincing England to betray Germany and convincing all remaining countries to agree to a draw—Italy gains control of the board. Italy's first deception is a plan with Austria to dismantle Turkey. Turkey believes Italy's initial assurance of non-aggression in 1901. Italy begins by excusing his initial silence due to a rough day at work, evoking empathy and *likability*. While they do not fall for subsequent lies, Turkey's initial gullibility cements Italy's first-strike advantage. Meanwhile, Italy proposes a long-term alliance with England against France, packaging several small truths with a big lie. The strategy succeeds,

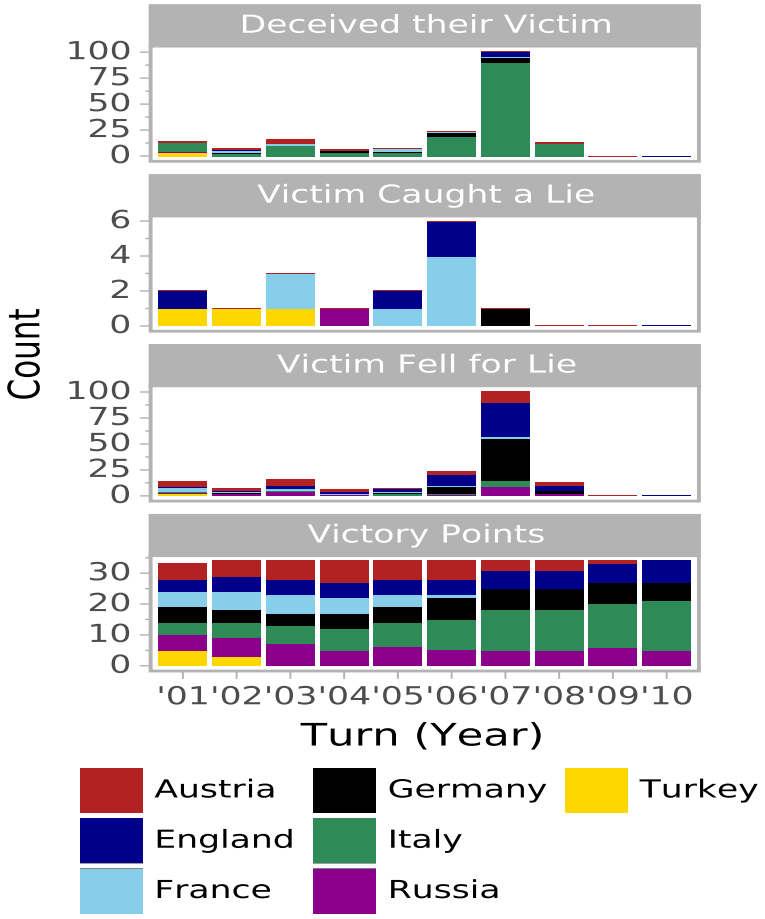

Figure 7.1: Counts from one game featuring an Italy (green) adept at lying but who does not fall for others' lies. The player's successful lies allow them to gain an advantage in points over the duration of the game. In 1906, Italy lies to England before breaking their relationship. In 1907, Italy lies to everybody else about wanting to agree to a draw, leading to the large spike in successful lies.

eliminating Italy's greatest threat.

Local threats eliminated, Italy turns to rivals on the other end of the map. Italy persuades England to double-cross its long-time ally Germany in a moment of *scarcity*: if you do not act now, there will be nowhere to expand. England accepts help from ascendant Italy, expecting *reciprocity*. However, Italy aggressively and successfully moves against England. The last year features a meta-game deception.

After Italy becomes too powerful to contain, the remaining four players team up. Ingeniously, Italy feigns acquiescence to a five-way draw, individually lying to each player and establishing *authority* while brokering the deal. Despite Italy's record of deception, the other players believe the proposal (annotating received messages from Italy as truthful) and expect a 1907 endgame, the year with the most lies. Italy goes on the offensive and knocks out Austria.

Each game has relationships that are forged and then riven. In another game, an honest attempt by a strong Austria to woo an ascendant Germany backfires, knocking Austria from the game. Germany builds trust with Austria through a believed fictional experience as a Boy Scout in Maine (*likability*). In a third game, two consecutive unfulfilled promises by an ambitious Russia leads to a quick demise, as their subsequent excuses and apologies are perceived as lies (failed *consistency*). In another game, England, France, and Russia simultaneously attack Germany after offering duplicitous assurances. Game outcomes vary despite the identical, balanced starting board, as different players use unique strategies to persuade, and occasionally deceive, their opponents.

7.2.2 Defining a lie

Statements can be incorrect for a host of reasons: ignorance, misunderstanding, omission, exaggeration. (Gokhman et al., 2012) highlight the difficulty of finding willful, honest, and skilled deception outside of short-term, artificial contexts (DePaulo et al., 2003). Crowdsourced and automatic datasets rely on simple nega-

tions (Pérez-Rosas et al., 2017) or completely implausible claims (e.g., "Tipper Gore was created in 1048" from (Thorne et al., 2018b)). While lawyers in depositions and users of dating sites will not willingly admit to their lies, the players of online games are more willing to revel in their deception.

We must first define what we mean by deception. Lying is a mischaracterization; it's thus no surprise that a definition may be divisive or the subject of academic debate (Gettier, 1963). We provide this definition to our users: "Typically, when [someone] lies [they] say what [they] know to be false in an attempt to deceive the listener" (Siegler, 1966). An orthodox definition requires the speaker to utter an explicit falsehood (Mahon, 2016); skilled liars can deceive with a patina of veracity. A similar definition is required for prosecution of perjury, leading to a paucity of convictions (Bogner et al., 1974). Indeed, when we ask participants what a lie looks like, they mention evasiveness, shorter messages, over-qualification, and creating false hypothetical scenarios (DePaulo et al., 2003).

7.2.3 Annotating truthfulness

Previous work on the language of Diplomacy (Niculae et al., 2015) lacks access to players' internal state and was limited to *post-hoc* analysis. We improve on this by designing our own interface that gathers players' intentions and perceptions in real-time (Section 7.4.1). As with other highly subjective phenomena like sarcasm (González-Ibáñez et al., 2011; Bamman and Smith, 2015), sentiment (Pang et al., 2008) and framing (Greene and Resnik, 2009), the intention to deceive is

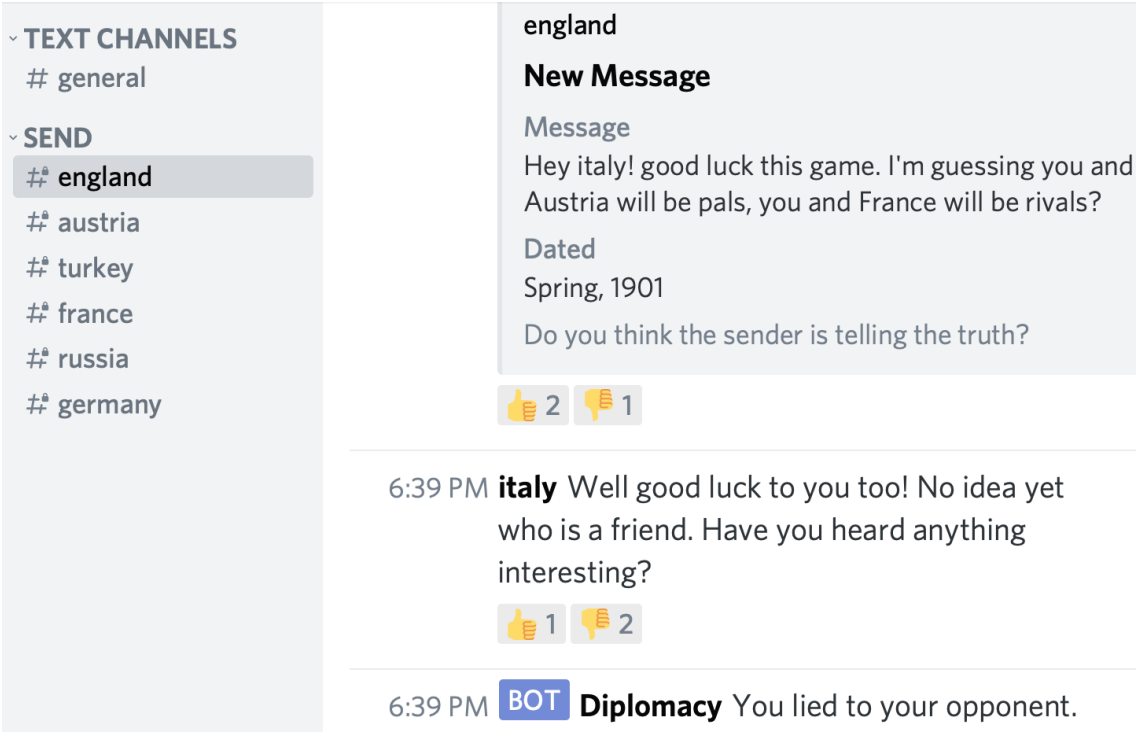

Figure 7.2: Every time they send a message, players say whether the message is truthful or intended to deceive. The receiver then labels whether incoming messages are a lie or not. Here Italy indicates they believe a message from England is truthful but that their reply is not.

reflective on someone's internal state. Having individuals provide their own labels for their internal state is essential as third party annotators could not accurately access it (Chang et al., 2020).

Most importantly, our gracious players have allowed this language data to be released in accordance with IRB authorized anonymization, encouraging further work on the strategic use of deception in long-lasting relations.[4]

[4]Data available at `http://go.umd.edu/diplomacy_data` and as part of ConvoKit `http://convokit.cornell.edu`.

7.3 Broader Applicability

This differs from previous work that does not follow the **expert-generated** paradigm. The most prominent past work on Diplomacy in the NLP community, (Niculae et al., 2015), **found** (Chapter 3) their data and thus could not release it to the public. This hampers follow-up applications of the research; a believable Diplomacy-playing (and speaking) bot cannot be trained if the raw language data is redacted and shuffled. We believe this work can set a paradigm for work outside of Diplomacy, and even NLP; the interface created for this project, as well as the pre and post-game user surveys can be modified for any conversational task (Chapter 8). Most importantly, building a relationship with data generators elevates the standard of the data and guarantees its liberal distribution. This mirrors the relationship with adaptation annotators (Chapter 6). Further work is necessary in codifying data standards—Show Your Data, not only your Work (Dodge et al., 2019).

7.4 Engaging a Community of Liars

This dataset requires both a social and technical setup: finding an online Diplomacy community and creating a framework for annotating messages between players.

7.4.1 Seamless Diplomacy Data Generation

We need two technical components for our study: a game engine and a chat system. We choose Backstabbr as an accessible game engine on desktop and mobile platforms: players input their moves and the site adjudicates game mechanics (Chiodini, 2020).[5] Our communication framework is atypical. Thus, we create a server on Discord, the group messaging platform most used for online gaming and by the online Diplomacy community (Coberly, 2019).[6] The app is reliable on both desktop and mobile devices, free, and does not limit access to messages. Instead of direct communication, players communicate with a bot; the bot does not forward messages to the recipient until the player annotates the messages (Figure 7.2). In addition, the bot scrapes the game state from Backstabbr to sync game and language data.

Annotation of lies is a forced binary choice in our experiment. We follow previous work that views linguistic deception as binary (Buller et al., 1996; Braun and Van Swol, 2016). However, explicitly calling a statement a lie is difficult, and people would prefer degrees of deception (Bavelas et al., 1990; Bell and DePaulo, 1996). Some studies make a more fine-grained distinction; for example, Swol et al. (2012) separate strategic omissions from blatant lies (we consider both deception). But, because we are asking the speakers themselves (and not trained annotators) to make the decision, we follow the advice from crowdsourcing to simplify the task as much as possible (Snow et al., 2008; Sabou et al., 2014). Long messages can contain both truths and lies, and we ask players to categorize these as lies since the truth

[5] `https://www.backstabbr.com`
[6] `https://www.discord.com`

can be a shroud for their aims.

7.4.2 Building a player base

The Diplomacy players maintain an active, vibrant community through real-life meetups and online play (Hill, 2014; Chiodini, 2020). We recruit top players alongside inexperienced but committed players in the interest of having a diverse pool.[7] Our experiments include top-ranked players and community leaders from online platforms, seasoned in-person tournament players with over 100 past games, and board game aficionados. These players serve as our foundation: during the initial design they helped us create a minimally annoying interface and a definition of a lie that would be consistent with Diplomacy play. Good players—as determined by active participation, annotation and game outcome—are asked to play in future games.

In traditional crowdsourcing tasks compensation is tied to piecework that takes seconds to complete (Buhrmester et al., 2011). Diplomacy games are different in that they can last a month. . . and people already play the game for free. Thus, we do not want compensation to interfere with what these players already do well: lying. Even the obituary of the game's inventor explains

> Diplomacy rewards all manner of mendacity: spying, lying, bribery, rumor mongering, psychological manipulation, outright intimidation, be-

[7] We recruit players from Diplomacy community forums, in-person tournaments, and board game clubs. We ask players if they are familiar with the rules of Diplomacy but do not have exclusionary qualification requirements; however, players that are not appropriately engaged are not invited to play further games, which happened in only a handful of cases.

Category	Value
Message Count	13,132
ACTUAL LIE Count	591
SUSPECTED LIE Count	566
Average # of Words	20.79

Table 7.2: Summary statistics for our train data (nine of twelve games). Messages are long and only five percent are lies, creating a class imbalance.

> trayal, vengeance and backstabbing (the use of actual cutlery is discouraged)" (Fox, 2013).

Thus, our goal is to have compensation mechanisms that get people to play this game as they normally would, finish their games, and put up with our (slightly) cumbersome interface. Part of the compensation is non-monetary: a game experience with players that are more engaged than the average online player.

To encourage complete games, most of the payment is conditioned on finishing a game, with rewards for doing well in the game. Players get at least $40 upon finishing a game.[8] Additionally, we provide bonuses for specific outcomes: $24 for winning the game (an evenly divisible amount that can be split among remaining players) and $10 for having the most successful lies, i.e., statements they marked as a lie that others believed.[9] Diplomacy usually ends with a handful of players dividing the board among themselves and agreeing to a tie. In the game described in Section 7.2.1, the remaining four players shared the winner's pool with Italy after 10 in-game years, and Italy won the prize for most successful lies.

[8]They receive $10 simply for having begun the study. No players dropped out of our games.

[9]The lie incentive is relatively small (compared to the $40 incentive for participation and up to $24 for winning) to discourage an opportunistic player from marking everything as a lie. Games were monitored in real-time and no player was found abusing the system (marking more than ∼20% lies).

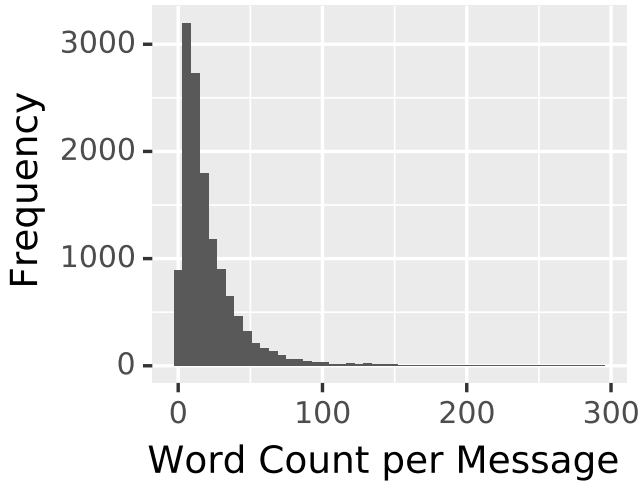

Figure 7.3: Individual messages can be quite long, wrapping deception in pleasantries and obfuscation.

7.4.3 Data overview

Table 7.2 quantitatively summarizes our data. Messages vary in length and can be paragraphs long (Figure 7.3). Close to five percent of all messages in the dataset are marked as lies and almost the same percentage (but not necessarily the same messages) are perceived as lies, consistent with the "veracity effect" (Levine et al., 1999). In the game discussed above, eight percent of messages are marked as lies by the sender and three percent of messages are perceived as lies by the recipient; however, the messages perceived as lies are rarely lies (Figure 7.4).

7.4.4 Demographics and self-assessment

We collect anonymous demographic information from our study participants: the average player identifies as male, between 20 and 35 years old, speaks English

		Receiver's perception	
		Truth	**Lie**
Sender's intention	**Truth**	**Straightforward** Salut! Just checking in, letting you know the embassy is open, and if you decide to move in a direction I might be able to get involved in, we can probably come to a reasonable arrangement on cooperation. Bonne journee!	**Cassandra** I don't care if we target T first or A first. I'll let you decide. But I want to work as your partner. ...I literally will not message anyone else until you and I have a plan. I want it to be clear to you that you're the ally I want.
	Lie	**Deceived** You, sir, are a terrific ally. This was more than you needed to do, but makes me feel like this is really a long term thing! Thank you.	**Caught** So, is it worth us having a discussion this turn? I sincerely wanted to work something out with you last turn, but I took silence to be an ominous sign.

Table 7.3: Examples of messages that were intended to be truthful or deceptive by the sender or receiver. Most messages occur in the top left quadrant (Straightforward). Figure 7.4 shows the full distribution. Both the intended and perceived properties of lies are of interest in our study.

as their primary language, and has played over fifty Diplomacy games.[10] Players self-assess their lying ability before the study. The average player views themselves as better than average at lying and average or better than average at perceiving lies.

In a post-game survey, players provide information on whom *they* betrayed and who betrayed *them* in a given game. This is a finer-grained determination than the *post hoc* analysis used in past work on Diplomacy (Niculae et al., 2015). We ask players to optionally provide linguistic cues to their lying and to summarize the game from their perspective.

[10]Our data skews 80% male and 95% of the players speak English as a primary language. Ages range from eighteen to sixty-four. Game experience is distributed across beginner, intermediate, and expert levels.

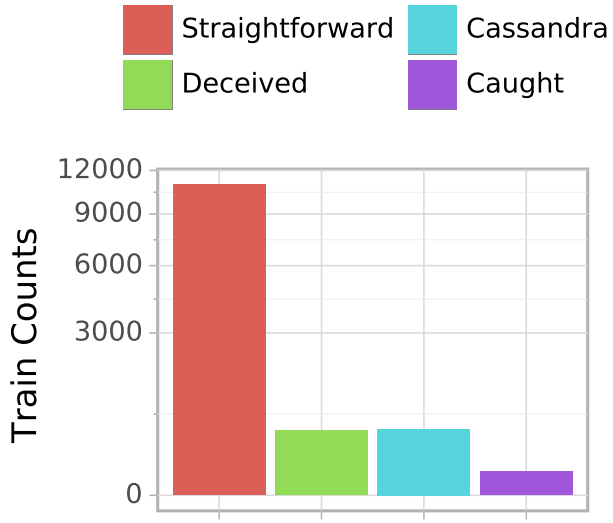

Figure 7.4: Most messages are truthful messages identified as the truth. Lies are often not caught. Table 7.3 provides an example from each quadrant.

7.4.5 An ontology of deception

Four possible combinations of deception and perception can arise from our data. The sender can be lying or telling the truth. Additionally, the receiver can perceive the message as deceptive or truthful. We name the possible outcomes for lies as Deceived or Caught, and the outcomes for truthful messages as Straightforward or Cassandra,[11] based on the receiver's annotation (examples in Table 7.3, distribution in Figure 7.4).

[11] In myth, Cassandra was cursed to utter true prophecies but never be believed. For a discussion of Cassandra's curse *vis a vis* personal and political oaths, see Torrance (2015).

7.5 Detecting Lies

We build computational models both to detect lies to better understand our dataset. The data from the user study provide a training corpus that maps language to annotations of truthfulness and deception. Our models progressively integrate information—conversational context and in-game power dynamics—to approach human parity in deception detection.

7.5.1 Metric and data splits

We investigate two phenomena: detecting what is *intended* as a lie and what is *perceived* as a lie. However, this is complicated because most statements are not lies: less than five percent of the messages are labeled as lies in both the ACTUAL LIE and the SUSPECTED LIE tasks (Table 7.2). Our results use a weighted F_1 feature across truth and lie prediction, as accuracy is an inflated metric given the class imbalance (Japkowicz and Stephen, 2002). We thus adopt an in-training approach (Zhou and Liu, 2005) where incorrect predictions of lies are penalized more than truthful statements. The relative penalty between the two classes is a hyper-parameter tuned on F_1.

Before we move to computational models for lie detection, we first establish the *human* baseline. We know when senders were lying and when receivers spotted a lie. Humans spot 88.3% of lies. However, given the class imbalance, this sounds better than it is. Following the suggestion of (Levine et al., 1999), we focus on the detection of lies, where humans have a 22.5 Lie F_1.

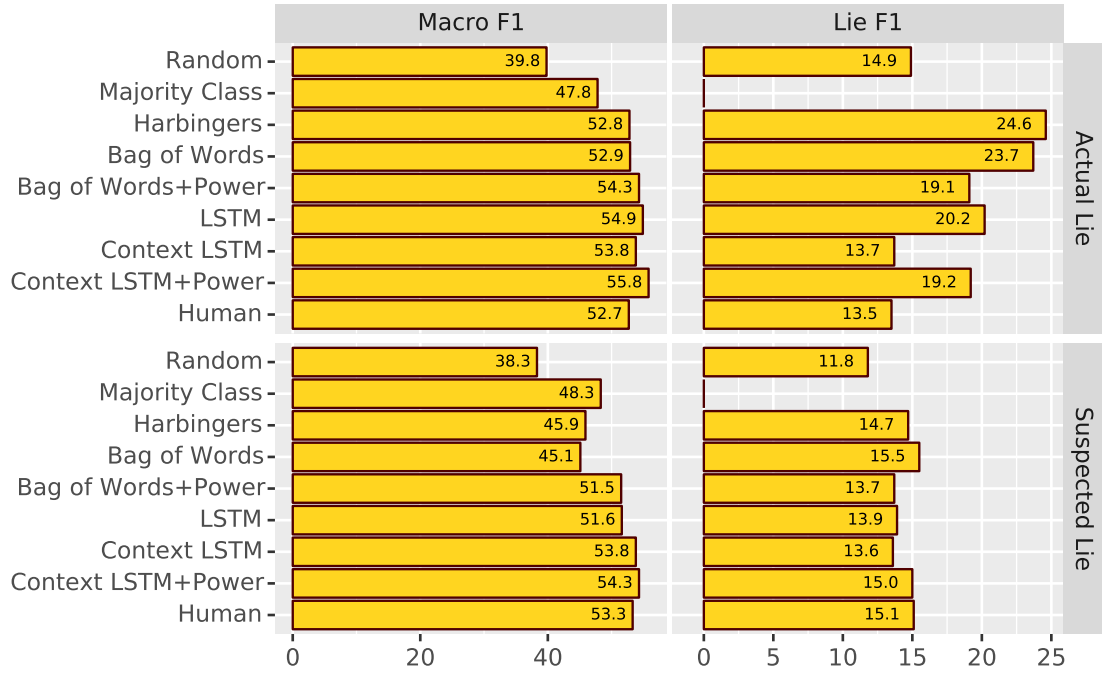

Figure 7.5: Test set results for both our ACTUAL LIE and SUSPECTED LIE tasks. We provide baseline (Random, Majority Class), logistic (language features, bag of words), and neural (combinations of a LSTM with BERT) models. The neural model that integrates past messages and power dynamics approaches human F_1 for ACTUAL LIE (top). For ACTUAL LIE, the human baseline is how often the receiver correctly detects senders' lies. The SUSPECTED LIE lacks such a baseline.

To prevent overfitting to specific games, nine games are used as training data, one is used for validation for tuning parameters, and two games are test data. Some players repeat between games.

7.5.2 Logistic regression

Logistic regression models (Background Section 2.3.1) have interpretable coefficients which show linguistic phenomena that correlate with lies. A *word* that occurs infrequently overall but often in lies, such as 'honest' and 'candidly', helps identify which messages are lies.

(Niculae et al., 2015) propose linguistic **Harbingers** that can predict deception. These are word lists that cover topics often used in interpersonal communication—*claims*, *subjectivity*, *premises*, *contingency*, *comparisons*, *expansion*, *temporal language associated with the future*, and *all other temporal language*. The Harbingers word lists do not provide full coverage, as they focus on specific rhetorical areas. A logistic regression model with all word types as features further improves F_1.

Power dynamics influence the language and flow of conversation (Danescu-Niculescu-Mizil et al., 2012, 2013; Prabhakaran et al., 2013). These dynamics may influence the likeliness of lying; a stronger player may feel empowered to lie to their neighbor. Recall that victory points (Section 7.2) encode how well a player is doing (more is better). We represent the power differential as the difference between the two players. Peers will have a zero differential, while more powerful players will have a positive differential with their interlocutor. The differential changes throughout the game, so this feature encodes the difference in the season the message was sent. For example, a message sent by an Italy with seven points to a Germany with two points in a given season would have a value of five.

7.5.3 Neural

While less interpretable, neural models are often more accurate than logistic regression ones (Ribeiro et al., 2016; Belinkov and Glass, 2019). We build a standard long short-term memory network (Hochreiter and Schmidhuber, 1997, LSTM in Section 2.3.4) to investigate if word sequences—ignored by logistic regression—can re-

veal lies.

Integrating message context and power dynamics improves on the neural baseline. A Hierarchical LSTM can help focus attention on specific phrases in long conversational contexts. In the same way it would be difficult for a human to determine *prima facie* if a statement is a lie without previous context, we posit that methods that operate at the level of a single message are limited in the types of cues they can extract. The hierarchical LSTM is given the context of previous messages when determining if a given message is a lie, which is akin to the labeling task humans do when annotating the data. The model does this by encoding a single message from the tokens, and then running a forward LSTM over all the messages. For each message, it looks at both the content and previous context to decide if the current message is a lie. Fine-tuning BERT (Devlin et al., 2019a) embeddings, introduced in Background Section 2.3.4, to this model did not lead to notable improvement in F_1, likely due to the relative small size of our training data. Last, we incorporate information about power imbalance into this model. This model approaches human performance in terms of F_1 score by combining content with conversational context and power imbalance.

7.6 Qualitative Analysis

This section examines specific messages where both players and machines are correctly identifying lies and when they make mistakes on our test set. Most messages are correctly predicted by both the model and players (2055 of 2475 messages);

		Model Prediction: Correct	Model Prediction: Wrong
Player Prediction	**Correct**	**Both Correct** Not sure what your plan is, but I might be able to support you to Munich.	**Player Correct** Don't believe Turkey, I said nothing of the sort. I imagine he's just trying to cause an upset between us.
	Wrong	**Model Correct** Long time no see. Sorry for the stab earlier. I think we should try to work together to stop france from winning; if we work together we can stop france from getting 3 more centers, and then we will all win in a 3, 4, or 5 way draw when the game is hard-capped at 1910.	**Both Wrong** I'm considering playing fairly aggressive against England and cutting them off at the pass in 1901, your support for that would be very helpful.

Table 7.4: An example of an ACTUAL LIE detected (or not) by both players and our best computational model (Context LSTM + Power) from each quadrant. Both the model and the human recipient are mostly correct overall (Both Correct), but they are both mostly wrong when it comes to specifically predicting lies (Both Wrong).

	Model Correct	**Model Wrong**
Player Correct	10	32
Player Wrong	28	137

Table 7.5: Conditioning on only lies, most messages are now identified incorrectly by both our best model (Context LSTM + Power) and players.

but this is because of the veracity effect. The picture is less rosy if we only look at messages the sender marks as ACTUAL LIE: both players and models are generally wrong (Table 7.5).

Both models and players can detect lies when liars get into specifics. In Diplomacy, users must agree to help one another through orders that stipulate "I will help another player move from X to Y". The in-game term for this is "support"; half the messages where players and computers correctly identify lies contain this word, but it rarely occurs in the other quadrants.

Models seem to be better at not falling for vague excuses or fantastical promises in the future. Players miss lies that promise long-term alliances, involve extensive apologies, or attribute motivation as coming from other countries' disinformation (*Model Correct*). Unlike our models, players have access to conversations with other players and accordingly players can detect lies that can easily be verified through conversations with other players (*Player Correct*).

However, ultimately most lies are believable and fool both models and players (*Both Wrong*). For example, all messages that contain the word "true" are predicted as truthful by both models and players. Many of these messages are relatively tame;[12] confirming the Pinocchio effect found by Swol et al. (2012). If liars can be detected when they wax prolix, perhaps the best way to avoid detection is to be terse and to the point.

Sometimes additional contextual information helps models improve over player predictions. For example, when France tells Austria "I am worried about a steamroller Russia Turkey alliance", the message is incorrectly perceived as truthful by both the player and the single-message model. However, once the model has context—a preceding question asking if Austria and Turkey were cooperating—it can detect the lie.

Finally, we investigate categories from the Harbingers (Niculae et al., 2015) word lists. Lies are more likely to contain *subjectivity* and *premises* while true messages include *expansion* phrases ("later", "additionally"). We also use specific

[12]Examples include "It's true—[Budapest] back to [Rumania] and [Serbia] on to [Albania] could position for more forward convoys without needing the rear fleet…" and "idk if it's true just letting u know since were allies".

words in the bag of words logistic regression model. The coefficient weights of words that express sincerity (e.g., "sincerely", "frankly") and apology (e.g., "accusation", "fallout", "alternatives") skew toward ACTUAL LIE prediction in the logistic regression model. More laid back appellations (e.g., "dude", "man") skew towards truthfulness, as do words associated with reconnaissance (e.g., "fyi","useful", "information") and time (e.g., "weekend", "morning"). Contested areas on the Diplomacy map, such as Budapest and Sevastopol, are more likely to be associated with lies, while more secure ones like Berlin, are more likely to be associated with truthful messages. These findings were not released to players during these data collection to avoid influencing players' language.

7.7 Related Work

Early computational deception work focuses on single utterances (Newman et al., 2003), especially for product reviews (Ott et al., 2012). But deception is intrinsically a discursive phenomenon and thus the context in which it appears is essential. Our platform provides an opportunity to observe deception in the context in which it arises: goal-oriented conversations around in-game objectives. Gathering data through an interactive game has a cheaper per-lie cost than hiring workers to write deceptive statements (Jurgens and Navigli, 2014).

Other conversational datasets are mostly based on games that involve deception including Werewolf (Girlea et al., 2016), Box of Lies (Soldner et al., 2019), and tailor-made games (Ho et al., 2017). However, these games assign individuals roles

that they maintain throughout the game (i.e., in a role that is supposed to deceive or in a role that is deceived). Thus, deception labels are coarse: an *individual* always lies or always tells the truth. In contrast, our platform better captures a more multi-faceted reality about human nature: everyone can lie or be truthful with everyone else, and they use both strategically. Hence, players must think about *every* player lying at any moment: "given the evidence, do I think this person is lying to me *now*?"

Deception data with conversational labels is also available through interviews (Pérez-Rosas et al., 2016), some of which allow for finer-grained deception spans (Levitan et al., 2018). Compared with game-sourced data, however, interviews provide shorter conversational context (often only a single exchange with a few follow-ups) and lack a strategic incentive—individuals lie because they are instructed to do so, not to strategically accomplish a larger goal. In Diplomacy, users have an intrinsic motivation to lie; they have entertainment-based and financial motivations to win the game. This leads to higher-quality, creative lies.

Real-world examples of lying include perjury (Louwerse et al., 2010), calumny (Fornaciari and Poesio, 2013), emails from malicious hackers (Dhamija et al., 2006), and surreptitious user recordings. But real-world data comes with real-world complications and privacy concerns. The artifice of Diplomacy allows us to gather pertinent language data with minimal risk and to access both sides of deception: intention and perception. Other avenues for less secure research include analyzing dating profiles for accuracy in self-presentation (Toma and Hancock, 2012) and classifying deceptive online spam (Ott et al., 2011).

Chapter 8: Quantity and (Mostly) Quality Through Hybridization[1]

As a dovetail between crowd-driven and expert-driven data sources, we propose a hybrid solution that pairs a **crowd-worker** *with* an **expert**. Our Diplomacy dataset (Chapter 7) shows that experts generate creative dialog. Our CANARD dataset (Chapter 5) shows that the crowd can perform simple tasks quickly, if there is a quality control process. We pair the two types of users together create dialog (Section 2.1.3) in an area more broadly applicable than Diplomacy: customer service. This creates a verisimilitude of a customer, simulated by a worker from the crowd, interacting with a customer service agent, simulated by an actual professional customer service agent. The resulting dataset illustrates the stark contrast in the language generated by anonymous crowd workers and experts. Furthermore, it demonstrates how NLP **generation** and **annotation** can be scaled through the crowd, while being quality controlled by an expert.

[1]Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang,Adel Youssef, and Mona Diab. Multi-domain goal-oriented dialogues(multidogo): Strategies toward curating and annotating large scale dialogue data. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4518–4528, 2019. Peskov planned and implemented some of the crowd-sourcing tasks, supervised the data collection thereof, wrote some of the task instructions, performed data analysis, and wrote most of the paper.

Role	Turn	Annotations
A	Hey there! Good morning. You're connected to LMT Airways. How may I help you?	DA = { elicitgoal }
C	Hi, I wonder if you can confirm my seat assignment on my flight tomorrow?	IC = { SeatAssignment }
A	Sure! I'd be glad to help you with that. May I know your last name please?	DA = { elicitslot }
C	My last name is Turker.	IC = { contentonly }, SL = {Name : Turker }
A	Alright Turker! Could you please share the booking confirmation number?	DA = { elicitslot }
C	I believe it's AMZ685.	IC = { contentonly }, SL = { Confirmation Number : AMZ685 }
...	...	...

Table 8.1: A segment of a dialog from the airline domain annotated at the turn level. This data is annotated with agent dialog acts (DA), customer intent classes (IC), and slot labels (SL). Roles C and A stand for "Customer" and "Agent".

8.1 The Goal of Creating Goal-Oriented Dialog

Modern Natural Language Understanding (NLU)—the integration of syntax, semantics, and inference (Winograd, 1972)—frameworks for dialog are data hungry. Processing goal-oriented dialog, which understands a user request and completes a related task with a clear goal within a limited number of dialog turns (Bordes et al., 2016), is an emblematic task of NLU: it requires extracting key information from free-form language. Large amounts of training data representative of the context is needed as human responses in goal-oriented dialogs are less predictable than those of automated systems (Bordes et al., 2016). For example, a broader context—like the questions in CANARD (Chapter 5)—is required to correctly interpret a command to "Please do this". This task can only be completed by seeing previous utterances, such as requests to book a flight on a specific day to a specific destination. A further

complication arises as multiple phrases can express a single intent depending on context: "book my flight", "finalize my reservation", "Yes, the 6 pm one" may all be refer to a flight-booking intent. Hence, we must generate entire *conversations*, rather than independent utterances.

Training goal-oriented dialog systems, and NLU in general, would benefit from large, varied, and ideally human-generated datasets. Joint-training and transfer learning (Dong et al., 2015; Devlin et al., 2019b) benefit natural language processing tasks; however, these approaches have yet to become widely used in dialog tasks due to a lack of large-scale datasets. Furthermore, end-to-end neural approaches benefit from such training data more than past work on goal-oriented dialog structured around slot filling (Lemon et al., 2006; Wang and Lemon, 2013).

Conveniently, the training data for goal-oriented dialogs occurs organically: people frequently converse with automated systems in customer service. Customers reach out to agents, automated bots or real people, to fulfill a domain-specific goal. The prevalence of human-machine interaction in customer service caused by personal virtual assistants and automated service portals has caused the amount of possible goals to multiply: ordering a meal, booking a plane ticket, and resolving an informational technology problem are all contexts in which goal-oriented dialog occurs. This creates an unbalanced conversation: agents operate within a set procedure and convey a patient and professional tone. In contrast, customers do not have this incentive; rather, they want to complete their task as quickly as possible. However, to date, the largest available multi-domain goal-oriented dialog dataset assigns similar dialog act annotations to both agents and customers (Budzianowski

et al., 2018).

We curate, annotate, and evaluate a large scale multi-domain set of goal oriented dialogs, `MultiDoGO`, to address the prior limitations. One way to simulate data—and not risk releasing personally identifying information—for a domain is to use a Wizard-of-Oz data gathering technique, which requires that participants in a conversation fulfill a role (Kelley, 1984). Popular goal-oriented datasets, DSTC (Williams et al., 2016) and `MultiWOZ` (Budzianowski et al., 2018) use this approach. Hence, our dataset is gathered from workers in the crowd paired with professional annotators using Wizard-of-Oz. The dataset generated comprises over 86K raw conversations of which 54,818 conversations are annotated at the turn level; this is a geometric increase over the number of utterances generated in Chapter 7. We investigate multiple levels of annotation granularity: annotating a subset of the data on both turn and sentence levels. Generating and annotating such data given its contextual setting is nontrivial. We furthermore illustrate the efficacy of our devised approaches and annotation decisions against intrinsic metrics and via extrinsic evaluation by applying neural baselines for **Dialog Acts**, **Intent Classification**, and **Slot Labeling**.

8.2 Existing Dialog Datasets

Chit-chat dialog without goals have been popular since ELIZA (Weizenbaum, 1966) and have been investigated with neural techniques (Li et al., 2016, 2017). However, these datasets cannot model goal-oriented tasks. Related dialog dataset

collections used for sequential question answering (Chapter 5) rely on dialog to answer questions, but the task differs from our use case of modeling goal-oriented conversations, hence leading to different evaluation considerations than downstream question answering (Choi et al., 2018; Reddy et al., 2019).

There are multiple existing goal-oriented dialog collections generated by humans through Wizard-of-Oz techniques (Kelley, 1984). The Dialog State Tracking Challenge, *aka* Dialog Systems Technology Challenge, (DSTC) spans eight iterations and entails the domains of bus timetables, restaurant reservations, and hotel bookings, travel, alarms, and movies (Williams et al., 2016). Frames (Asri et al., 2017) has 1369 dialogs about vacation packages. `MultiWOZ` contains 10,438 dialogs about Cambridge hotels and restaurants (Budzianowski et al., 2018). Some dialog datasets specialize in a single domain. In addition to the datasets mentioned in Background Section 2.1.3, ATIS (Hemphill et al., 1990) comprises speech data about airlines structured around formal airline flight tables. Similarly, the Google Airlines dataset purportedly contains 400,000 templated dialogues about airline reservations (Wei et al., 2018).[2]

8.3 `MultiDoGO` Dataset Generation

Generating and **annotating** a dataset of this scale requires task design, data collection, and post-task quality control.

[2]The Google Airlines dataset has not been released to date despite the existence of a paper describing it.

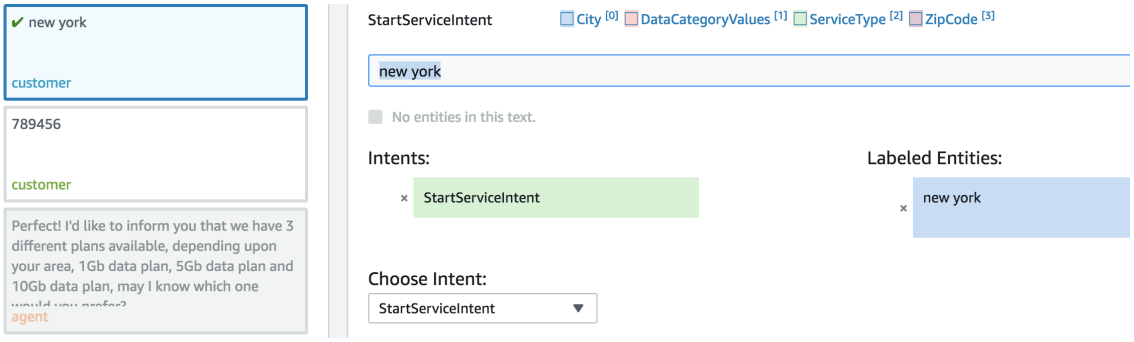

Figure 8.1: Crowd sourced annotators select an intent and choose a slot in our custom-built Mechanical Turk interface. Entire conversations are provided for reference. Detailed instructions are provided to users, but are not included in this figure. Options are unique per domain.

8.3.1 Defining Dialog

We define the dialog terminology that is discussed in our design process. A turn is a sequence of speech/text sentences by a participant in a conversation. A sentence is a period delimited sequence of words in a turn. A turn may comprise multiple sentences. We do use the term utterance to refer to a unit (turn or sentence, spoken or written by a participant).[3]

In our devised annotation strategy, we distinguish between dialog speech acts for agents vs. customers. In `MultiDoGO`, the agents' speech acts [DA] are annotated with generic class labels common across all domains, while customer speech acts are labeled with intent classes [IC]. Moreover, we annotate customer utterances with a further level of detail.

[3]We acknowledge that the term utterance is controversial in the literature (Pareti and Lando, 2018)

8.3.2 Data Collection Procedure

We employ professional annotators, who we train, and crowd-sourced workers from Mechanical Turk (MTurkers) to generate conversational data using a Wizard-of-Oz approach.[4] In each conversation, the data associates assumes the role of an agent while the MTurkers act as customers. In an effort to source competent MTurkers, we require that each MTurker have a Human Intelligence Task (HIT) accuracy minimum of 90%, a location in the United States, and have completed HITs in the past.[5] We give each agent a prompt listing the supported request types (dialog acts) and pieces of information (slots) needed to complete each request to structure goal-oriented conversations between the customer and agent. We also specify criteria such as minimal conversation length, number of goals, and number of complex requests to increase conversation diversity (Figure 8.2). We explicitly request that neither agents nor customers use any personally identifiable information.[6] At an implementation level, we create a custom, web interface for the MTurkers and data associates that displays our instructions next to the current dialog. This allows each participant to quickly refer to our instructions without stopping the conversation. `MultiDoGO` follows a familiar Wizard-of-Oz elicitation procedure and curates data for multiple domains akin to previous data collection efforts such as `MultiWOZ`. However, `MultiDoGO` comprises more varied domains, is a magnitude larger, and is

[4]The professional annotators are salaried employees of the company engaging in this research. They were staffed on this project full-time for three months. Training sessions were conducted in person for a full day explaining the annotation guidelines and answering any questions that arose.

[5]Qualified MTurkers were allowed to complete the generation and annotation tasks multiple times.

[6]They are however encouraged to fabricate information for slots. (e.g., John Smith as a name.

curated with prompts to ensure diverse conversations.

This is a novel collection strategy as we explicitly guide/prod the participants in a dialog to engage in conversations with specific biases such as intent change, slot change, multi-intent, multiple slot values, slot overfilling and slot deletion. For example, in the Fast Food domain, participants pretend that they were ordering fast food from a drive-thru. After making their initial order, they were instructed to change their mind about what they were ordering: "I'd like a burger. No wait, can you make that a chicken sandwich?". In the Financial domain, we asked participants request multiple intents such as "I'd like to find my routing number and check my balance."[7] To that end, our collection procedure deliberately attempts to guide the dialog flow to ensure diversity in dialog policies.

8.4 Data Annotation

Annotation classifies the thousands of conversations in our dataset. Of particular interest, a direct comparison of using experts versus the crowd is made in Section 8.4.2. Our annotators use a web interface (Figure 8.1) to select the appropriate intent class for an utterance out of a list of provided options. They use their cursors to highlight slot value character spans within an utterance and then select the corresponding slot label from a list of options to annotate slot labels. The output of this slot labeling process is a list of ⟨slot-label, slot-value, span⟩ triplets for each utterance.

[7]For a full list of conversational biases with examples, please see the Appendix.

8.4.1 Annotated Dialog Tasks

Our dataset has three types of annotation: agent dialog acts [DA], customer intent classes [IC], and slot labels [SL]. We intentionally decouple agent and customer speech act tags into the categories DA and IC to produce more fine-grained speech act tags than past iterations of dialog datasets. Intuitively, agent DAs are consistent across domains and more general in nature, since agents have a standard form of response. On the other hand, customer ICs are domain-specific and can entail reserving a hotel room or ordering a burger, depending on the domain. A conversation example with annotations is provided in Table 8.1.

Agent Dialog Acts (DA) Agent dialog acts are the most straightforward of our annotation tasks. There are eight possible DAs in all domains: ElicitGoal, ElicitSlot, ConfirmGoal, ConfirmSlot, EndGoal, Pleasantries, Other. Elicit Goal/Slot indicates that the agent is gathering information. Confirm Goal/Slot indicates that the agent is confirming previously provided information. The EndGoal and Pleasantries tags, identify non-task related actions.[8] Other indicates that the selected utterance was not one of the other possible tags. Agent dialog acts are consistent across domains and are often abstract (e.g., ElicitIntent, ConfirmSlot).

Customer Intent Classes (IC): Unlike agent DA, customer IC vary for each domain and are more concrete. For example, the Airline domain has a "BookFlight" IC, Fast Food has an "OrderMeal" IC, and Insurance has an "OrderPolicy" IC in our annotation schema. Customer intents can overlap across domains (e.g.,

[8]EndGoal is a frequently occurring case of Pleasantry when the agent informs the customer that the goal has been completed and asks if anything else is required.

OpeningGreeting, ClosingGreeting) and other times be domain specific (e.g., RequestCreditLimitIncrease, OrderBurger, BookFlight).

Slot Labels (SL): Slot labeling is a task contingent on customer intent Classes. Certain intents require that additional information, namely slot values, be captured. For instance, to open a bank account, one must solicit the customer's social security number. Slots can overlap across intents (e.g., Name, SSN Number) or they can be unique to a domain-specific intent (e.g., CarPolicy).

8.4.2 Annotation Design Decisions

Decoupled Agents and Customers Label Sets Agents and customers have notably different goals and styles of communication. However, past dialog datasets do not make this distinction at speech act schema level. Specificity is important for generating unique customer requests, but a relatively formulaic approach is required of agents across different industries. Our distinction between the customer and agent roles creates training data for a bot that explicitly simulates agents.

Annotation Unit Granularity: Sentence vs. Turn Level An important decision, which is often under-discussed, is the proper semantic unit of text to annotate in a dialog. Commonly, datasets provide annotations at the turn level (Budzianowski et al., 2018; Asri et al., 2017; Mihail et al., 2017). However, turn level annotations can introduce confusion for IC datasets, given multiple intents may be present in different sentences of a single turn. For instance, consider the turn, "I would like to book a flight to San Francisco. Also, I want to cancel a

Dialog Act	Intent Classes	Slot Labels
0.701	0.728	0.695

Table 8.2: Inter Source Annotation Agreement (ISAA) scores quantifying the agreement of crowd sourced and professional annotations.

flight to Austin." Here, the first sentence has the BookFlight intent and the second sentence has the CancelFlight intent. A turn level annotation of this utterance would yield the multi-class intent (BookFlight, CancelFlight). In contrast, a sentence level annotation of this utterance identifies that the first sentence corresponds to BookFlight while the second corresponds to CancelFlight. We annotate a subset our data—2,500 conversations per domain for 15,000 conversations in total—at the sentence as well as turn level to assess the design choice on downstream accuracy (Table 8.8). The remainder of our dataset is annotated only at the turn level.

Professional vs. Crowd-Sourced Workers for Annotation For annotation, we compare and contrast professional annotators to crowd sourced annotators on a subset of data. Professional annotators assign DA, IC, and SL tags to the 15,000 conversations annotated at both the turn and sentence level; statistics for these conversations are given in Table 8.7. In an effort to decrease annotation cost, we employ crowd source annotators via Mechanical Turk to label an additional 54,818 conversations rated as Good or Excellent quality during data collection.[9] We provide statistics for this set of crowd annotated data in Table 8.3. To compare the quality of crowd sourced annotations against professional annotations, we use both strategies to annotate a shared subset of 8,450 conversations. We devise an

[9]Users are still paid for conversations that are not evaluated as such. Since this happened after conversation generation, we do not need to exclude users responsible for bad conversations from future ones.

Domain	Elicited	Good/Excellent	IC/SL	DA/IC/SL
Airline	15100	14205	7598	6287
Fast Food	9639	8674	7712	4507
Finance	8814	8160	8002	6704
Insurance	14262	13400	7799	7434
Media	33321	32231	19877	12891
Software	5562	4924	3830	2753
Total	**86698**	**81594**	**54818**	**40576**

Table 8.3: Total number of conversations per domain: raw conversations Elicited; Good/Excellent is the total number of conversations rated as such by the agent annotators; (IC/SL) is the number of conversations annotated for Intent Classes and Slot Labels only; (DA/IC/SL) is the total number of conversations annotated for Dialog Acts, Intent Classes, and Slot Labels.

Inter Source Annotation Agreement (ISAA) metric to measure the agreement of these crowd sourced and professionally sourced annotations. ISAA is a relaxation of Cohen κ, intended to count partial agreement of multi-tag labels. ISAA defines two sets of tags, A and B, to be in agreement if there is at least one "shared" tag in both A and B. A and B reflect the majority labels agreed upon per source (professionals or crowd workers). We report ISAA for the DA, IC, and SL tasks in Table 8.2. Crowd sourced and professional annotations have substantial overlap between their annotations. Therefore, the crowd can be used for **annotation** for NLP tasks, if the annotations are verified to be comparable to experts.

8.4.3 Quality Control

Three processes enforce data quality. During data collection, our **experts** report on the quality of each conversation. Specifically, the expert grades the conversation on a scale from "Unusable", "Poor", "Good", to "Excellent". They follow instructions around coherence, whether the dialog achieved the purported goal, etc.,

Bias	Airlines	Fast Food	Finance	Insurance	Media	Software
IntentChange		1443				
MultiIntent	2200	1913	1799	1061	607	2295
MultiValue		354				
Overfill			1486	2763		
SlotChange	4207	2011	2506	3321	570	2085
SlotDeletion		333				
Total	6407	6054	5791	7145	1177	4380

Table 8.4: Number of conversations per domain collected with specific biases. Fast Food had the maximum number of biases. MultiIntent and SlotChange are the most used biases.

to decide on the chosen rating. We keep conversations with “Good" or “Excellent" ratings in subsequent annotation to maximize the quality of our dataset.

Second, each conversation is annotated at least twice. We resolve inconsistent annotations by selecting the annotation given by the majority of annotators for an item.[10] We calculate inter-annotator agreement with Fleiss' κ and find “substantial agreement”.[11] Our annotators must pass a qualification test as well as maintain an on-going level of accuracy in randomly distributed test questions throughout their annotation. Third, we pre-process our data to remove issues, such as duplicate conversations and improperly entered slot value spans. Further pre-processing details are in Section 8.5.

8.4.4 Dataset Characterization and Statistics

The `MultiDoGO` dataset is the most diverse dialog dataset due to covering more domains and being generated, rather than scraped from existing and dubiously reliable data sources (e.g., Ubuntu forums). Table 8.3 shows the statistics

[10] We drop annotations in which there is no agreement.

[11] We use Fleiss' κ unlike in the earlier profession/crowd worker comparison as we have more than two annotators for this task.

Metric	DSTC 2	WOZ2.0	M2M	MULTIWOZ	MULTIDOGO
Number of Dialogs	1,612	600	1,500	8,438	40,576
Total Number of Turns	23,354	4,472	14,796	115,424	813,834
Total Number of Tokens	199,431	50,264	121,977	1,520,970	9,901,235
Avg. Turns per Dialog	14.49	7.45	9.86	15.91	20.06
Avg. Tokens Per Turn	8.54	11.24	8.24	13.18	12.16
Total Unique Tokens	986	2,142	1,008	24,071	70,003
Number of Unique Slots	8	4	14	25	73
Number of Slot Values	212	99	138	4,510	55,816
Number of Domains	1	1	1	7	6
Number of Tasks	1	1	2	2	3

Table 8.5: `MultiDoGO` is several times larger in nearly every dimension to the pertinent datasets as selected by Budzianowski et al. (2018). We provide counts for the training data, except for FRAMES, which does not have splits. Our number of unique tokens and slots can be attributed to us not relying on carrier phrases.

for `MultiDoGO` raw conversations generated, rated as Excellent or Good, and annotated for DA, IC and SL. Table 8.4 shows the number of conversations per domain reflecting the specific biases used.

`MultiDoGO` is several orders of magnitude larger than comparable datasets as reflected in nearly every dimension: the number of conversations, the length of the conversation, the number of domains, and the diversity of the utterances used. Table 8.5 provides comparative statistics.

Domain	#Conv	#Turn	#Turn/Conv	#Sentence	#Intent	#Slot
Airline	2,500	39,616	15.8 (15)	66,368	11	15
Fast Food	2,500	46,246	18.5 (18)	73,305	14	10
Finance	2,500	46,001	18.4 (18)	70,828	18	15
Insurance	2,500	41,220	16.5 (16)	67,657	10	9
Media	2,500	35,291	14.1 (14)	65,029	16	16
Software	2,500	40,093	16.0 (15)	70,268	16	15

Table 8.6: Data statistics by domain. Conversation length is in *average (median)* number of turns per conversation. Inter-annotator agreement (IAA) is measured with Fleiss' κ for the three annotation tasks: Agent DA (DA), Customer IC (IC), and Slot Labeling (SL).

We provide summary statistics for the subset of our data annotated at both turn and sentence granularity in Table 8.7. This describes the total size of the

Domain	Turn-level IAA	Sentence-level IAA
Airline	0.514/0.808/0.802	0.670/0.788/0.771
Fast Food	0.314/0.700/0.624	0.598/0.725/0.607
Finance	0.521/0.827/0.772	0.700/0.735/0.714
Insurance	0.521/0.862/0.848	0.703/0.821/0.826
Media	0.499/0.812/0.725	0.678/0.802/0.758
Software	0.508/0.748/0.745	0.709/0.764/0.698

Table 8.7: Inter-annotator agreement (IAA) is measured with Fleiss' κ for the three annotation tasks: Agent DA (DA), Customer IC (IC), and Slot Labeling (SL).

		Airline			Fast Food			Finance		
Model	Annot	DA	IC	SL	DA	IC	SL	DA	IC	SL
MFC	S	60.57	33.69	38.71	57.14	25.42	61.92	51.73	37.37	34.07
LSTM	S	97.20	90.84	74.16	90.40	86.09	72.93	93.90	90.06	69.09
ELMO	S	**97.32**	**91.88**	**86.55**	**91.03**	**87.95**	**77.51**	**94.07**	**91.15**	**77.36**
MFC	T	33.04	32.79	37.73	33.07	25.33	61.84	36.52	38.16	34.31
LSTM	T	**84.25**	89.15	75.78	**66.41**	87.35	73.57	76.19	92.30	70.92
ELMO	T	84.04	**89.99**	**85.64**	65.69	**88.96**	**79.63**	**76.29**	**94.50**	**79.47**
		Insurance			Media			Software		
Model	Annot	DA	IC	SL	DA	IC	SL	DA	IC	SL
MFC	S	56.87	38.37	53.75	57.02	30.42	82.06	58.14	33.32	53.96
LSTM	S	**94.73**	93.30	75.27	**94.27**	92.35	90.84	93.22	90.95	69.48
ELMO	S	94.63	**94.27**	**88.45**	**94.27**	**93.32**	**93.99**	**93.66**	**92.25**	**76.04**
MFC	T	36.39	39.42	54.66	29.90	31.82	78.83	36.79	33.78	54.84
LSTM	T	**75.37**	94.75	76.84	**77.94**	94.35	87.33	**83.32**	89.78	72.34
ELMO	T	75.34	**95.39**	**89.51**	77.81	**94.76**	**91.48**	82.97	**90.85**	**76.48**

Table 8.8: Dialog act (DA), intent class (IC), and slot labeling (SL) F1 scores by domain for the majority class, LSTM, and ELMobaselines on data annotated at the sentence (S) and turn (T) level. Bold text denotes the model architecture with the best performance for a given annotation granularity, i.e., sentence or turn level. Red highlight denotes the model with the best performance on a given task across annotation granularities.

data per domain in number of conversations, turns, the unique number of intents and slots, and inter-annotator agreement (IAA) for both turn and sentence level annotations. DA annotations have much higher IAA in sentence-level annotations compared to turn-level annotation, most notably in the Fast Food domain. IC and SL annotations reflect a slightly higher IAA in Turn level annotation granularity compared to Sentence level.

Agent Instructions

Imagine you work at a bank. Customers may contact you about the following set of issues: checking account balances (checking or savings), transferring money between accounts, and closing accounts.

GOAL: Answer the customer's question(s) and complete their request(s).

For any request, you will need to collect at least the following information to be able to identify the customer: name, account PIN *or* last 4 digits of SSN.

For giving information on balances, or for closing accounts, you will also need the last 4 digits of the account number.

For transferring money, you will also need: last 4 digits of account to move from, last 4 digits of account to move to, and the sum of money to be transferred.

Your customer may ask you to do only one thing; that's okay, but make sure you confirm you achieved everything the Customer wanted before completing the conversation. Don't forget to signal the end of the conversation (see General guidelines)

Figure 8.2: Agents are provided with explicit fulfillment instructions. These are quick-reference instructions for the Finance domain. Agents serve as one level of quality control by evaluating a conversation between Excellent and Unusable.

	Airline		Fast Food		Finance		Insurance		Media		Software	
A	Single	Joint	Single	Joint	Single	Joint	Single	Joint	Single	Joint	Single	Joint
S	97.32	**97.44**	91.03	**91.26**	94.07	**94.27**	94.63	**94.99**	94.27	**94.47**	93.66	**94.00**
T	84.04	**84.64**	**65.69**	65.35	**76.29**	75.68	75.34	**75.89**	77.81	**78.56**	82.97	**83.76**

Table 8.9: Joint training of ELMo on all agent DA data leads to a slight increase in test performance. However, we expect stronger joint models that use transfer learning should see a larger improvement. Bold text denotes the training strategy, i.e., single domain (Base) or multi-domain (Joint), with the best performance for a given annotation granularity. Red highlight denotes the strategy with the highest DA F1 score across annotation granularities.

8.5 Dialog Classification Baselines

We pre-process, create dataset splits, and evaluate the performance of three baseline models for each domain on `MultiDoGO`.

Pre-processing: We pre-process the corpus of dialogs for each domain to

remove duplicate conversations and utterances with inconsistent annotations. The most common source of inconsistent annotations in our dataset is imprecise selection of slot label spans by annotators, which results in sub-token slot labels. While much of this inconsistent data could likely be recovered by mapping each character span to the nearest token span, we drop these utterances to ensure these errors have no effect on our experimental results. Our post-processed data is pruned to approximately 90% of the original size. We form splits for each domain at the conversation level by randomly assigning 70% of conversations to train, 10% to development, and 20% to test. Conversation level splits enable the application of contextual models to our dataset, as each conversation is assigned to a single split. However, our conversation level splits result in imbalanced intent and slot label distributions.

Models: We evaluate the performance of two neural models on each domain. The first is a bi-directional LSTM (Hochreiter and Schmidhuber, 1997) with GloVe word embeddings, a hidden state of size 512, and two fully connected output layers for slot labels and intent classes. The second model, ELMo, resembles LSTM architecture but it additionally uses pre-trained ELMo (Peters et al., 2018) embeddings in addition to GloVe word embeddings, which are kept frozen during training (Sections 2.3.4). We concatenate these ELMo and GloVe embeddings. As a sanity check, we also include a most frequent class (MFC) baseline. The MFC baseline assigns the most frequent class label in the training split to every utterance u' in the test split for both DA and IC tasks. To adapt the MFC baseline to SL, we compute the most frequent slot label $\text{MFC}(w)$ for each word type w in the training set. Then given a test utterance u', we assign the pre-computed, most frequent slot $\text{MFC}(w')$ to each

word $w' \in u'$ if w' is present in the training set. If a given word $w' \in u'$ is not present the training set, we assign the *other* slot label, which denotes the absence of a slot, to w'. We use AllenNLP (Gardner et al., 2017) for models and metrics. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001 to train the LSTM and ELMo models for 50 epochs, using batch sizes 256 and 128. In addition, we use early stopping on the validation loss with a tolerance of 10 epochs to prevent over-fitting.

Evaluation Metrics: We report micro F1 score to evaluate DA and IC. We use a span based F1 score, implemented in the seqeval library, to evaluate SL.[12]

8.5.1 Results

DA/IC/SL Results. Table 8.8 presents the MFC, LSTM, and ELMo results for each domain, on the subset of 15,000 conversations annotated at both the turn and sentence levels. LSTM, and ELMo outperform MFC across all domains at the turn and sentence level. ELMo obtains a modest increase in IC accuracy of 0.41 to 2.20 F1 points and a significant increase in SL F1 score on all domains over the LSTM baseline. Concretely, ELMo boosts SL F1 performance by 3.16 to 13.17 F1 points. We see the biggest SL gains on the Insurance domain, where sentence level ELMo has a 13.17 point F1 gain and turn level ELMo has a 12.67 point F1 gain. ELMo increases sentence and turn level SL F1 scores by 12.38 and 9.86 F1 points for the airlines domain. Both LSTM and ELMo yield similar F1 scores on DA classification for which the difference in performance of these models is within one F1 point across

[12] https://github.com/chakki-works/seqeval

all domains. The Fast Food domain yields the overall lowest absolute F1 scores. Recall that Fast Food had the most diverse dialogs (biases) as per Table 8.4 and the lowest IAA as per Table 8.7.

Sentence vs. Turn Level Annotation Units. Turn level annotations increase the difficulty of the DA classification task in our LSTM and ELMo results. This finding is evidenced by DA accuracy of our models on the Fast Food domain, for which F1 score is up to 25 F1 points lower for turn level annotations than sentence level annotations. We believe the increased difficulty of turn level DA is driven by a corresponding increase in the ambiguity of turn level dialog acts. This assertion of greater turn level DA ambiguity is supported by the lower inter annotator agreement (IAA) scores on turn level DA, which range from 0.314 to 0.521, than the IAA scores for sentence level DA, which range from 0.598 to 0.709. This experimental result highlights the importance of collecting sentence level annotations for conversational DA datasets. Somewhat surprisingly, our models have similar IC F1 and SL F1 scores on turn and sentence level annotations. We posit that the choice of annotation unit has a lesser impact on the IC and SL tasks because customer utterances are more likely to focus on a single speech act, whereas Agent utterances may be more complex in comparison and include a greater number of speech acts.

Joint Training on Agent DA. Agent DA classification naturally lends itself to joint training, given agent DAs are shared among all domains. To explore the benefits of multi-domain training, we jointly train an agent DA classification model on all domains and report test results for each domain separately. These results are provided in Table 8.9. This straightforward technique leads to a consistent but

less than one point improvement in F1 scores. We expect that more sophisticated transfer learning methods (Liu et al., 2017; Howard and Ruder, 2018) could generate larger improvements for these domains.

Overall, there is room for improvement, especially for the SL task, across all domains. Consequently, `MultiDoGO` should be a relevant benchmark for developing new state-of-the-art NLU models for the foreseeable future.

8.6 Conclusion

We present `MultiDoGO`, a new Wizard-of-Oz dialog dataset that is the largest human-generated, multi-domain corpora of conversations to date. The scale and range of this data provides a test-bed for future work in joint training and transfer learning. Moreover, our comparison of sentence and turn level annotations provides insight into the effect of annotation granularity on downstream model performance.

The data collection and annotation methodology used to gather `MultiDoGO` can efficiently scale across languages. Several pilot experiments aimed at collecting Spanish dialogs in the same domains have shown preliminary success in quality assessment. The production of a NLU dataset with parallel data in multiple languages would be a boon to the cross-lingual research community. To date, cross-lingual NLU research (Upadhyay et al., 2018; Schuster et al., 2018) has relied on much smaller parallel corpora.

By pairing **crowd**-sourced labor (Chapter 5) with **experts** (Chapters 6 and 7), we ensure quality and diversity in **generated** conversations while scaling to multiple

domains and tasks. We show that by adopting a modular annotation strategy, the crowds can reliably **annotate** dialogs at a level commensurate with trained professional annotators. Without the expert, our data would be just as large, but it could not be trusted.

There is a stark difference in quality of the **generated** language between the crowd-sourced workers and the experts, in this case in-house customer service agents. The crowd-sourced workers have a financial incentive to complete the task as quickly as possible and contribute sentences that are occasionally prosaic, ungrammatical, or repeated.[13] In our case, these incentives mimic those of the usual customer and does not undermine the realism of the conversation. But, should datasets be *large* or should they be *accurate* in future work where these incentives are not desirable? We conclude with areas for future work that must balance quantity and accuracy to be successful (Chapter 9).

[13]We pay workers for completing a full valid conversation; conversely, our agents converse with the workers as long as necessary and are not paid for each conversation.

Chapter 9: Conclusions on Natural Language Processing Data

In this thesis, we create natural language processing datasets using three types users: **crowd-workers**, **experts**, and a **hybrid** combination. We argue that improving data quality with reliable data **generators** and **annotators** is paramount towards establishing new NLP tasks. As examples, we propose a new task, cultural adaptation, that uses verified cultural experts for the creation of gold labels (Chapter 6). Additionally, we introduce a novel self-annotated deception dataset by working with top players from the Diplomacy community (Chapter 7). Last, we create the largest goal-oriented dialogue dataset by pairing Amazon customer support associates with crowd workers (Chapter 8).

These datasets could not be **found** or **crowd-sourced**.[1] Several projects show the limitations of creating large datasets in that way. Using text-to-speech to automatically generate questions scales at the expense of diversity and realism in the data (Chapter 3). Using an **expert** to design, but not generate, a formulaic dataset for assessing coreference resolution creates unlikely phrases (Chapter 4). Using the **crowd** to generate question rewrites can increase the amount of training data for question answering, but requires extensive quality control (Chapter 5).

[1] Or at least be comparable in quality if they were.

Two independent directions for future work both use experts to create new datasets. First, Diplomacy2.0 extends our work on Diplomacy (Section 9.1). Second, the World Trade Organization and the Federal Reserve are two large organizations whose data can be annotated by experts to create legal corpora (Section 9.2) We conclude with a rationalization for upfront investment in data (Section 9.3).

9.1 Hybridization of Diplomacy: Diplomacy2.0

We **generate** and **annotate** Diplomacy data to study deception (Chapter 7). We can improve this existing dataset through further annotation of dialog acts (Section 9.1.1). This level of annotation would allow us to build a bot that can communicate in Diplomacy. We want to merge communication with actions, and need game experts to map these dialog acts to game moves (Section 9.1.2). Further studies with the Diplomacy community will create a new dataset, Diplomacy2.0, to study strategic interactions between computers and humans.

9.1.1 Data for Communication

In our initial Diplomacy work, players use deception for multiple purposes: lying to convey plans and forge alliances ("Let's team up to take out Germany"); lying about past actions ("My computer was acting up... I didn't make the move I wanted"); and lying to build relationships ("you live in Maine? I went to a boy scout camp there!"). We want to formally identify and annotate these categories as we want our bot to use rhetorical strategy: using empathy (Sedoc et al., 2020),

emotional intensity (Mohammad, 2018), and hedging (Islam et al., 2020). Creating an ontology of dialog acts is typical for new domains—telephone conversations (Stolcke et al., 2000), scientific articles (Teufel and Moens, 2002), or political speeches (Thomas et al., 2006). We will need Diplomacy experts to identify this new ontology. Since we are interested in general dialog acts, rather than capturing the nuance of Diplomacy, we can use the **crowd** for the actual annotation.[2] Thus this will be a hybrid approach of using **experts** for design and the **crowd** for scaling akin to Chapter 8.

We can build a bot that devises a game strategy and *communicates* their intention to other players through these dialog acts. Communications assume a player p creating a message directed at recipient q. These messages can concern a third player r and can have one of three modes: **d**eclarative (I am asserting something is true), **i**nterrogative (I am querying your beliefs), and **p**ropositional (I am asking you to do something). Each message is parameterized by the mode m, actions a, and time t. In addition to machine-readable fields, each message allows for arbitrary free text; this can contain elaboration, hedging, or motivation. A bot will need to have training examples of modes to **generate** an effective message.

9.1.2 Data for Action

Most actions correspond to the orders that players can submit in a game of Diplomacy: moving a unit, supporting another unit, building a new unit. In

[2]We may have to source the crowd from Diplomacy players if the generalist annotation proves inaccurate.

addition to these actions, which are *explicit* in Diplomacy, we consider *implicit* actions in Diplomacy negotiations: an alliance between two players (pursue the same goals, support whenever possible), a non-aggression pact (no explicit cooperation but no attacks), and betrayal (break either an alliance or non-aggression pact to hurt a player's position). Thus, an agent can pledge to do a future action $p(a, t+1, m = \text{d})$ (here and below we use the first letter of the modes), ask the recipient to do something $q(a, t+1, m = \text{p})$; communicate about a third player's intentions; $r(a, t, m = \text{i})$; propose an alliance $p(\text{ally}, q, m = \text{p})$; ask whether a third player is allied with the recipient $r(\text{ally}, q, m = \text{i})$; propose a betrayal of player r at time t, etc. After receiving a message, a recipient can confirm that the message is consistent with their knowledge, reject the message as inconsistent with their knowledge, ignore the message, or reply with a counter-offer.

Purely strategic data exists in other larger datasets (Paquette et al., 2019; Bakhtin et al., 2021). While we can train from self-play, that would ignore previous games of Diplomacy (Niculae et al., 2015), tutorials on how to play Diplomacy (e.g., opening strategies for each of the countries),[3] and commentary on Diplomacy games.[4] We will use these corpora to bootstrap the strategy engine as Bakhtin et al. (2021) show that Diplomacy systems trained from scratch do not converge to the same behavior as human players. Vetting the computer's moves and combining them with our past press will require Diplomacy experts, as automatic mapping would be noisy; for example, Turkey moving into the Black Sea may map to "my

[3] `https://diplomacyopenings.wordpress.com/`
[4] `https://youtu.be/b4GHbg5--Ag?t=138`

computer was acting up", rather than a strategic message.

With this training, our Diplomacy2.0 agent will construct a per-game knowledge base. Upon receiving a message, it will be added to the board state as an entry in the knowledge base in addition to annotations of whether the agent believes the message was true or not. For games with humans, we will also ask the sender to provide ground-truth annotation of whether the message was true for *post hoc* refinement of our bots.

9.1.3 Evaluation Through Human Studies

The key challenge is to train this engine to work with other players, something not yet attempted in AI for Diplomacy. Given a game state, Diplomacy2.0 must produce coherent, useful messages to communicate with other players, and evaluate whether the messages the agent received are truthful. Like past work (Chapters 6 and 7), there is no existing gold label to evaluate this communication. Hence, we will define the evaluation for this task.

While our overall goal is for Diplomacy2.0 to win games against other agents, we also want to have evaluations specific to its subcomponents, like the hypothetical coreference pipeline in ContraCAT (Chapter 4). This will include generating messages, correctly inferring opponents' stance, and cooperative play.

Generating Messages To evaluate whether Diplomacy2.0 can generate messages, we will evaluate precision and recall of generated messages given a fixed board state compared to human-generated messages in our annotated corpus of dialog acts.

While we do not assume that humans are optimal, this is a useful sanity check of our communications: if they are consistent with human gameplay, it suggests Diplomacy2.0 is generating messages that are relevant and consistent with the game state. We will also compare against precision and recall for deception detection compared to our previous techniques.

Correctly Inferring Opponent Stance A key theoretical component of negotiations is successfully predicting what opponents want and will do. To evaluate whether Diplomacy2.0 can do this, we will predict (given a board state and messages annotated for dialog acts) what each of the players will do next. Given a game history, we can compute both standard precision and recall metrics of predicted actions compared to the true history and mean reciprocal rank of the historical actions compared to a ranked list of predictions. This will show the bot's understanding of game actions.

Cooperative Play To bridge between no-press Diplomacy and our communication-enabled setting, we will evaluate how well the bot can do without communication by self-play against copies of itself and other bots incapable of speech. If they are better able to coordinate and win against their foes than a mute bot, then our bot is able to communicate. Next, we will evaluate the complete Diplomacy2.0 system. Finally, we will evaluate AI agents in mixed environments with humans and AI agents. Here, the evaluation is more nuanced: our goal is not just whether agents can win games. AI agents must be able to cooperate with both humans and computers, correctly predict betrayals, *a*nd ultimately win the game.

9.2 Understanding Organizations with Economic and Legal Experts

We can use experts for new purposes in addition to extending our past work with experts. As one possibility, expert annotation can provide insights on organizational decision making at scale by facilitating the analysis of deliberative processes. Major organizations often engage in formal dialogues prior to making decisions. Many of these discussions are recorded for posterity. But, reading through dozens of years of proceedings and legal documents is onerous for a person; however, the task is trivial for a computer. We propose to apply NLP techniques to solving real-life problems in the public domain. Specifically, we will annotate then quantitatively analyze publicly available historical text data from major organizations to identify patterns of decision-making. Two major organizations, the World Trade Organization and the Federal Reserve Board, can be analyzed with the same abstract methodology.

Data is integral to machine learning; models are only as useful if the underlying training data is relevant to the desired task and correctly labeled. Apropos **annotation** in the social sciences, Text as Data (Grimmer and Stewart, 2013) reviews the strengths and limitations of computation for political science, noting that, "Ambiguities in language, limited attention of coders, and nuanced concepts make the reliable classification of documents difficult." This future work will require annotation to be consistent across projects—the terms and phrases used in our two organizations will differ but we propose creating a shared abstract ontology that can be used for any organizational dialogue—and even across disciplines, ideally defining

a gold standard for guideline creation and quality control in the process.

The task required: identifying ideas in dense legal documents will be particularly challenging due to the nuance of the task. Even defining what constitutes an idea is a challenge. An idea can be the subject of conversation (Grimmer, 2010), a meme (Leskovec et al., 2009), a narrative (Oates, 2014), etc. Therefore, defining a universal annotation schema, writing unambiguous guidelines, and finding skilled users to perform this annotation will be no small feat. Fortunately, the adaptation (Chapter 6) and Diplomacy (Chapter 7) datasets required experts and bespoke instructions and `MultiDoGO` (Chapter 8) required a universal annotation schema. We will use this newly created representation of the data to answer domain-specific questions. The influence of ideas can be identified with NLP techniques (Zhang et al., 2016). We propose to formalize this task by, once-again (Section 9.1.1), creating an ontology for organizational decision-making. Then, we will be able to analyze and understand our data as a sequence of *ideas* rather than mere text.

9.2.1 The World Trade Organization

World Trade Organization (WTO) cases, meticulously documented in hundreds of pages, are a strong fit for this research question as they occur over long-time periods, include arguments made by third party countries, and are publicly available (Busch and Pelc, 2019). We will answer a concrete social science problem: can final case outcomes be predicted based on the involved parties and their submitted arguments? This prediction has international economic implications, as trade dis-

putes set precedents outside of the goods in question: a ruling on tires may affect trade in bananas through legal precedent.

Working with domain experts is necessary to solve this research question due to the complex legal language of the WTO. For background, disputants and third parties submit opinions on legal precedents that should be considered in shaping the final judgment. Expertise will entail the legal education needed to identify key WTO cases and relevant precedents in documents, which can be over a hundred pages long. A successful annotation of these submitted opinions—both literal references and their abstract ideas—in chronological legal documents would identify how trade decisions are made at the highest level of international governance and create one of the largest datasets of legal data to date. This in turn could be used to set a gold standard training dataset for legal NLP that can be used for predicting future case outcomes.

9.2.2 The Federal Reserve Board

The Federal Reserve Board decides monetary policy in the United States and provides public documentation of their proceedings for the past 40 years. Panel participants state their opinion at the beginning and the end of the session and cast a vote. The social science question we aim to answer is if the opinion of participants of the panel changes during a session in line with the the common knowledge effect (Gigone and Hastie, 1993). Psychology research identifies that priors are the principal driver of decision-making in a group, but it has never been

able to verify the effect at scale. In addition to answering this question, the presence of a binary vote will allow us to study the language of dissent and conformity. We will see if majority voters are more likely to change their mind during the panel discussion than minority ones. The dense economic language of these panels will require expert annotators with a background understanding of economics and finance.

9.3 Creating Timeless Natural Language Processing Datasets

The above two projects will require large-scale collaboration with the appropriate subject matter experts for extended periods of time. However, datasets that have withstood the test of time in natural language processing were also painstakingly created and quality controlled. The Penn Treebank (Marcus et al., 1993) was collected and refined for years using graduate students in linguistics as annotators. The annotation process had extensive experimental design, annotators underwent extensive training, and the data was evaluated for disagreements. That effort caused graduate students today to learn about it.

The granularity and quantity of NLP datasets continues to increase as machine learning expands to new languages and tasks. Quality control is usually an afterthought in a conference paper paradigm that rewards quantity. However, this mindset introduces room for error, potentially with real-life repercussions (Wallace et al., 2021). The importance of NLP to modern day life in communication, information gathering, and commerce means that decisions made in an academic context can have wide-ranging implications. Authoritative, realistic, and diverse datasets

are less likely to contain errors or artifacts and more likely to be used in years to come than larger datasets derived from Wikipedia or crowd-sourced knowledge.

Recent work questions conventional wisdom about data in NLP. Rodriguez et al. (2021) question the paradigm of using quantitative leaderboards in question answering, given the disparity of question difficulties. van der Goot (2021) question the paradigm of using a development set for model tuning. Kummerfeld (2021) question the qualification requirements for Mechanical Turk workers. Last, Karpinska et al. (2021) question the output of Mechanical Turk workers for evaluation. Pillutla et al. (2021) create a divergence metric to compare artificial and human language data. The common thread between these open questions is that they address data and not models. We show that working with domain experts and focusing on data quality can address complicated natural language processing challenges.

Appendix A: Adaptation

Our adaptation (Chapter 6) appendix contains our entire human-collected dataset and a sample of our computational approaches for adaptation.

Table A.1 shows German→American Veale NOC items. Table A.2 shows American→German Veale NOC items. Table A.3 shows German→American Veale NOC items. Table A.4 shows American→German Veale NOC items.

Table A.5 shows our WikiData predictions, Table A.6 shows our 3CosAdd predictions. and Table A.7 shows our Learned Adaptations predictions. We pose several background questions about Wikipedia and WikiData as well:

A.1 Wikipedia Q&A

Are the Wikipedia pages in German and English visited from the associated country? Yes; the Wikipedias for the respective languages are most used by visitors located in those countries: 63% of German wikipedia was visited from Germany and 32% of English Wikipedia was visited from the United States in the past year.[1]

Are the top Wikipedia topics notably different across languages? Yes; less than a quarter of top 500 searches for 2019 are identical across English and German.

[1] https://stats.wikimedia.org/

Does WikiData cover areas outside of the United States? *Wikipedia* coverage does not mean that *WikiData* annotations are conducted equally across German and American entities. Analyzing WikiData[2] reveals a discrepancy in coverage of Germans and Americans.

Out of 8,126,559 titles, 1,030,762 include a reference to the United States in any capacity. However, only 184,692 contain a reference to (broader) Germany. This imbalance is significant but has enough German items for our methodology. As WikiData is a maintained resource, there is room for future additional coverage and standardization of fields.

Countries use different names throughout history. While the United States of America is straightforward, Germany includes several variations, such as: German Empire, the Kingdom of Bavaria, the Kingdom of Prussia, etc. The WikiData feature-based approach can be used for other countries as well (. . . or anything that is consistently coded). For example, there are 65,957 Russian, 152,701 French, and 48,026 Chinese items in WikiData.[3]

Are the top Wikipedia topics necessarily belonging to the culture? No; the top 10 most visited German Wikipedia includes a cultural potpurri: Germany, Greta Thurnberg, Asperger Syndrome, Game of Thrones, and Freddie Mercury. While there are *uniquely* German entities in the longer list—ZDF, Capital Bra, The Cratez, Niki Lauda—we **cannot** conclude that all top entities in a language belong culturally to a given country. Therefore, we need a stricter methodology.

[2] we use a full 1.2 Terabyte dump as of 10.26.20

[3] the modern day name countries only

Where does one find entities? We rely on a human-sourced dataset: Veale's Non-Official Characterization list (Veale, 2016). This list contains 1031 people, real and fictional, such as Daniel Day-Lewis, Anton Chekhov, and Bridget Jones. These people are annotated with properties, one of which is conveniently their address. There are 25 people with a German location and 575 with an American one. Removing fictional characters written by non-nationals causes the German leaves the list with 20 entities. An American author filters the list of Americans down to 35 iconic ones with achievements that span politics, music, activism, athletics, and pop culture.

Wikipedia provides another avenue for gauging popular topics in a language. We manually filter the top 500 German/English Wikipedia topics to remove non-German/non-American entities; Game of Thrones and Unix-Shell are popular in the German Wikipedia, but they are not culturally idiosyncratic. For the 2019 German Wikipedia we are left with roughly 200 items, which we further reduce down to 120 after putting a cap on pop culture entities. For the American counterpart, over 300 items are culturally American. We add a three-year filter to remove pop items to make it comparable to the German one.

A.2 Data

Entity	**Human Adaptation: NOC German→American**
Adolf Eichmann	Andrew Jackson, Andrew Jackson, Franklin D. Roosevelt, Nathan Bedford Forrest, Steve Bannon
Angela Merkel	Barack Obama, Donald Trump, Hillary Clinton, Hillary Clinton, Hillary Clinton, Joe Biden
Baron Munchausen	Captain America, Daniel Bolger, Joseph Smith, Paul Bunyan, Robert Jordan , Yankee Doodle
Carl von Clausewitz	Alfred Thayer Mahan, Dwight D. Eisenhower, Henry Knox, Robert E. Lee, Ulysses S. Grant
Friedrich Nietzsche	Ayn Rand, Henry David Thoreau, Henry Thoreau, Jordan Peterson, William James
Henry Kissinger	Henry Kissinger, Henry Kissinger, John Kerry, Madeleine Albright, Richard Nixon
Immanuel Kant	Benjamin Franklin, John Dewey, John Locke, John Rawls, Robert Nozick
Johann Sebastian Bach	Aaron Copland, Elvis Presley, Elvis Presley, Irving Berlin, Johnny Cash, Scott Joplin
Johann Wolfgang von Goethe	Edgar Allan Poe, Ernest Hemingway, Walt Whitman
Johannes Gutenberg	Benjamin Franklin, Bill Gates, Eli Whitney, Thomas Edison
Joseph Goebbels	David Duke, Franklin D. Roosevelt, George Rockwell, Rupert Murdoch, david duke
Karl Lagerfeld	Anna Wintour, Anna Wintour, Marc Jacobs, Ralph Lauren, Ralph Lauren, Ralph Lauren, Ralph Lauren
Karl Marx	Angela Davis, Beck, Bernie Sanders, John Jay, John Rawls, John Rawls
Leni Riefenstahl	DW Griffeth, David Wark Griffith, Frank Capra, Judy Garland
Ludwig van Beethoven	Aaron Copland, Aaron Copland, Aaron Copland, Elvis Presley, Frank Sinatra, George Gershwin, George Gershwin, Scott Joplin
Marlene Dietrich	Bette Davis, Clara Bow, Elizabeth Taylor, Marilyn Monroe, William Tecumseh Sherman
Martin Luther	Barry Goldwater, Brigham Young, Joseph Smith, Joseph Smith, Joseph Smith
Otto von Bismarck	Abraham Lincoln, George Washington, George Washington, Ulysses S. Grant
Pope Benedict XVI	Billy Graham, Billy Graham, Brigham Young, John Carroll , Seán Patrick O'Malley
Richard Wagner	Charles Ives, Frank Sinatra, Leonard Bernstein, Philip Glass

Table A.1: Veale NOC German→American adaptations.

Entity	Human Adaptation: NOC American→German adaptations
Abraham Lincoln	Helmut Kohl, Konrad Adenauer, Wilhelm Friedrich Ludwig von Preußen, Willy Brandt, Willy Brandt
Al Capone	Adolf Leib, Carlos Lehder-Rivas, Jan Marsalek, Nasser Abou-Chaker, Nasser About-Chaker
Alfred Hitchcock	Bernd Eichinger, Bernd Eichinger, Michael Bully Herbig, Roland Emmerich, Wim Wenders
Benedict Arnold	Hansjoachim Tiedge, Otto von Bismarck, Otto von Bismarck, Robert Blum
Bill Gates	Andreas von Bechtolsheim, Carl Benz, Dietmar Hopp, Konrad Zuse
Britney Spears	Helene Fischer, Herbert Grönemeyer, Jeanette Biedermann, Nena, Til Schweiger
Charles Lindbergh	Ferdinand von Richthofen, Heinrich Horstman, Karl Wilhelm Otto Lilienthal, Ludwig Hofmann, Wernher von Braun
Donald Trump	Adolf Hitler, Adolf Hitler, Carsten Maschmeyer, Christian Lindner
Elvis Presley	Peter Kraus, Rammstein, The Scorpions, Udo Lindenberg, Udo Lindenberg
Ernest Hemingway	Günter Grass, Hermann Hesse, Johann Wolfgang von Goethe, Karl May, Martin Walser
Frank Lloyd Wright	Gerhard Richter, Hugo Häring, Karl Lagerfeld, Max Dudler, Walter Gropius
George Washington	Friedrich II, Heinrich I, Konrad Adenauer, Otto I. der Große, Otto von Bismarck
Henry Ford	Carl Benz, Carl Benz, Carl Benz, Ferdinand Porsche, Gottlieb Wilhelm Daimler
Hillary Clinton	Angela Merkel, Angela Merkel, Angela Merkel, Kramp-Karrenbauer, Sahra Wagenknecht
Homer Simpson	Alf, Heidi, Pumuckl, Werner, Werner - Beinhart!
Jack The Ripper	Armin Meiwes, Der Bulle von Tölz, Joachim Kroll, Karl Denke, Rudolf Pleil
Jay Z	Capital Bra, Marteria, Sido, Sido, Sido
Jimi Hendrix	Bela B., Gisbert zu Knyphausen, Herbert Grönemeyer, Rudolf Schenker, Spider Murphy Gang
John F. Kennedy	Hanns Martin Schleyer, Willy Brandt, Willy Brandt, Wolfgang Schäuble
Kim Kardashian	Carmen Geiss, Gina-Lisa Lohfink, Heidi Klum, Heidi Klum, Sarah Connor
Louis Armstrong	Günter Sommer, Helmut Brandt, Jan Delay, Michael Abene, Mozart
Marilyn Monroe	Heidi Klum, Ingrid Steeger, Marlene Dietrich, Micaela Schäfer, Uschi Glas
Michael Jordan	Dirk Nowitzki, Dirk Nowitzki, Dirk Nowitzki, Franz Beckenbauer, Michael Schuhmacher

Neil Armstrong	Alexander Gerst, Sigmund Jähn, Sigmund Jähn, Ulf Merbold, Wernher von Braun
Noam Chomsky	Helmut Glück, Juergen Habermas, Jürgen Habermas, Ludwig Wittgenstein, Wilhelm Röntgen
Oprah Winfrey	Anne Will, Arabella Kiesbauer, Maybrit Illner, Thomas Gottschalk, Thomas Gottschalk
Orville Wright	Carl Benz, Gustav Otto, Gustav Weißkopf, Otto Lilienthal, Wernher von Braun
Richard Nixon	Franz Josef Strauss, Helmut Kohl, Ludwig Erhard, Ludwig Erhard, Richard von Weizsäcker
Rosa Parks	Anne Wizorek, Marie Juchacz, Sophie Scholl, Sophie Scholl, Vera Lengsfeld
Serena Williams	Andrea Petkovic, Boris Becker, Sabine Lisicki, Steffi Graf, boris becker
Steve Jobs	Carl Benz, Dietmar Hopp, Dietmar Hopp, Karl Lagerfeld
Steven Spielberg	Michael Bully Herbig, Roland Emmerich, Roland Emmerich, Roland Emmerich, Wim Wenders
Superman	Bibi Blocksberg, Fix and Foxi, Maverick, Superman, Till Eulenspiegel
Tiger Woods	Boris Becker, Martin Kaymer, Martin Kaymer, Michael Schumacher, Serge Gnabry
Walt Disney	Axel Springer, Christian Becker, Franz Mack, Gerhard Hahn, Rötger Feldmann

Table A.2: Veale NOC American→German adaptations.

Entity	Human Adaptation: Wikipedia German→American
ARD	NPR, PBS, PBS
Adolf Hitler	Donald Trump, Donald Trump, Franklin D. Roosevelt, Franklin D. Roosevelt, Franklin D. Roosevelt
Airbus	Boeing, Boeing, Boeing, Boeing, Lockheed Martin
Albert Einstein	Carl Sagan, J. Robert Oppenheimer, J. Robert Oppenheimer, John Forbes Nash Jr., Thomas Edison
Alice Merton	Ariana Grande, Elle King, K.T. Tunstall, P!NK, Vanessa Carlton
Alternative für Deutschland	Libertarian Party , Republican Party, Tea Party movement
Andrea Nahles	Elizabeth Warren, Hillary Clinton, Nancy Pelosi, Tammy Duckworth
Andrej Mangold	Kawhi Leonard, Kevin Durant, Kris Humphries, Yao Ming
Annalena Baerbock	Al Gore, Al Gore, Alexandria Ocasio-Cortez, Bernie Sanders, Jill Stein
Anne Frank	Anna Green Winslow, Clara Barton, Emmett Till, Kunta Kinte
Annegret Kramp-Karrenbauer	Condoleezza Rice, Hillary Clinton
AnnenMayKantereit	Guns N' Roses, Milky Chance, Polar Bear Club, Red Hot Chili Peppers
Apache 207	Fetty Wap, Tekashi 69, XXXTentacion, Zayn Malik
Arnold Schwarzenegger	Chuck Norris, Dwayne Johnson, Ronnie Coleman, Sylvester Stallone, Sylvester Stallone
BMW	Cadillac, Cadillac, Chevrolet, Chrysler
Babylon Berlin	Game of Thrones, Man From U.N.C.L.E., Peaky Blinders , The Americans, Turn
Baden-Württemberg	California, Chicago metropolitan area, San Diego, Southern United States, Texas
Bastian Yotta	Chad Johnson, Colton Underwood, Dan Bilzerian
Bauhaus	Frank Lloyd Wright
Bayerischer Rundfunk	NPR, National Public Radio, National Public Radio, national public ra
Bayern	Florida, New York, The Confederacy
Benjamin Piwko	Bruce Lee, Colton Underwood, Derek Hough
Berlin	New York City, Portland Oregon, Washington D.C., Washington D.C., Washington D.C.
Berliner Mauer	Border Patrol Police, Mason–Dixon line, Mason–Dixon line, US-Mexican border
Bertolt Brecht	Tennessee Williams, Tennessee Williams
Björn Höcke	Lindsey Graham, Mike Pence
Borussia Dortmund	Golden State Warriors, New England Patriots, New England Patriots

Brandenburg	Maryland, New York, Northeastern United States, Richmond Virginia, Virginia
Bruno Ganz	Clint Eastwood, Ethan Hawke, Marlon Brando, Robert De Niro, Robert De Niro
Bundespräsident	First Lady, President of the United States, Speaker of the House
Bundeswehr	Department of Defense , US military, United States Armed Forces, United States Army
Capital Bra	Drake, Eminem, Eminem, Kanye West, Kendrick Lamar
Carola Rackete	American Civil Liberties Union, Dawn Wooten, Rosa Parks, Whale Wars
Carolin Kebekus	Amy Schumer, Sarah Silverman, Tina Fey, Tina Fey
Charité	Call the Midwife, Grey's Anatomy, Grey's Anatomy, The Queen's Gambit
Chris Töpperwien	Gordon Ramsey , Guy Fieri, Jeff Probst
Christoph Waltz	Anthony Hopkins, Christoph Waltz, Denzel Washington
Dark	Stranger Things, Stranger Things
Deutsche Bahn	Amtrack, Norfolk Southern Railway, Union Pacific Corporation
Deutsche Demokratische Republik	Confederate States of America, Confederate States of America, Texas, The Confederacy, The Confederate States of America
Deutsche Nationalhymne	Born in the U.S.A., Lazy Eye , Star Spangled Banner, The Star Spangled Banner
Deutschland	America, America, Continental United States, USA, United States, United States
Dieter Bohlen	Billy Joel, Blake Shelton, Daryl Hall, Paula Abdul, Ryan Seacrest
Dirk Nowitzki	LeBron James, Michael Jordan, Shaquille O'Neal
Doreen Dietel	Jessica Alba, Lisa Kudrow, Warrick Brown
Dreißigjähriger Krieg	American Civil War, American Civil War, American Indian Wars, Civil war
Elisabeth von Österreich-Ungarn	Edith Roosevelt, Hillary Clinton, Jackie Kennedy
Elyas M'Barek	Adam Sandler, Adam Sandler, Chris Pine
Europawahl in Deutschland 2019	2018 United States elections, American presidential election 2020, Us election 2018
Europäisches Parlament	North Atlantic Council, Representative of the United States of America to the European Union, United Nations, United States Congress
Evelyn Burdecki	Hannah Brown, Kaitlyn Bristowe, Kim Kardashian, Kim Kardashian
FC Bayern München	Dallas Cowboys, Dc United, New York Yankees, New York Yankees, New York Yankees
Falco	David Bowie, Frederick William Schneider III, MC Hammer, Michael Jackson

Ferdinand Sauerbruch	Ben Carson, Ben Carson, Cornelius P. Rhoads, Jonas Salk, Virginia Apgar
Flughafen Berlin Brandenburg	Cincinnati Subway, DCA , John F. Kennedy International Airport, LaGuardia Airport
Frankfurt am Main	Chicago, Los Angeles, Los Angeles, New York City, Washington D.C.
Fritz Honka	Ted Bundy, Ted Bundy, Ted Bundy, Zodiac
Hamburg	Chicago, Chicago, Los Angeles, New York, Philadelphia
Hannelore Elsner	Elizabeth Taylor, Jane Lynch, Julia Roberts
Heidi Klum	Chrissy Teigen, Cindy Crawford, Gigi Hadid, Karlie Kloss, Tyra Banks
Heinz-Christian Strache	Anthony Weiner, Ben Carson, Donald J. Trump, Rob Ford, Roger Stone
Helene Fischer	Beyoncé, Kelly Clarkson, Taylor Swift, Taylor Swift
Hessen	Arizona, Illinois, Mid-Atlantic , Napa County California
Holocaust	Chattel Slavery, Japanese interned in American camps, Slavery in the United States
Ich bin ein Star – Holt mich hier raus!	Survivor, Survivor
Jürgen Klopp	Bill Belichick, Bill Belichick, John Wooden
Kevin Kühnert	Bernie Sanders, Bernie Sanders, Bernie Sanders, Pete Buttigieg
Klaus Kinski	Christopher Lee, Clark Gable, John Wayne, Robert Pattinson, Robert Pattinson
Kontra K	50 Cent, Eminem, Eminem, Jesus Is King, Travis Scott
Köln	Boston, Chicago, Chicago, Houston
Leila Lowfire	Paris Hilton, Sasha Grey, Zendaya
Leipzig	Denver, Detroit, Miami, San Diego
Lena Meyer-Landrut	Ariana Grande, Kelly Clarkson, Kelly Clarkson, Meghan Trainor, Selena Gomez
Liechtenstein	Connecticut, Mexico, Philippines, Victoria British Columbia
Lisa Martinek	Julie Benz, Katherine Heigl, Mandy Moore, Meryl Streep
Ludwig van Beethoven	Aaron Copland, Aaron Copland, Aaron Copland, Aaron Copland, Elvis Presley, Frank Sinatra, George Gershwin, George Gershwin, Scott Joplin
Lufthansa	Delta, United, United Airlines, United Airlines
Luxemburg	Canada, Connecticut, Mexico, Victoria British Columbia
Mark Forster	Bruno Mars, Post Malone
Mero	DaBaby, Fetty Wap, Lil Nas X, Lil Nas X, Post Malone
Michael Schumacher	Dale Earnhardt, Dale Earnhardt, James Gordon, Jeff Gordon, Tiger Woods

München	Chicago, Los Angeles, New York City, New York City, Washington D.C.
Nico Santos	Harry Styles, Justin Bieber, Shawn Mendes
Niki Lauda	Dale Earnhardt, Dale Earnhardt Jr., Jeff Gordon, Jeff Gordon, Tiger Woods
Norddeutscher Rundfunk	NPR, NPR, National Public Radio, PBS, Sirius XM
Nordrhein-Westfalen	California, California
Philipp Amthor	Alexandria Ocasio-Cortez, Ben Shapiro
RAF Camora	Bad Bunny, Drake, Drake , Eminem, Future
Rammstein	Green Day, Metallica, Metallica, Metallica, Sum 41
Rhein	Mississippi, Mississippi River, Mississippi River
Robert Habeck	Al Gore, Bernie Sanders, Jill Stein, Ralph Nader
Rudi Assauer	Dave Roberts, Gregg Berhalter, Tom Flores, Vince Lombardi, Vince Lombardi
Sahra Wagenknecht	Alexandria Ocasio-Cortez, Elizabeth Warren, Elizabeth Warren, Elizabeth Warren, Nancy Pelosi
Sarah Connor	Beyoncé, Britney Spears, Mariah Carey
Schweiz	Canada, Canada, Iowa, Mexico, United States
Sebastian Kurz	Alexandria Ocasio-Cortez, Greg Abbott, Justin Trudeau, Justin Trudeau, Mitch McConnell
Serge Gnabry	Clint Dempsey, JuJu Smith-Schuster, Phillip Rivers, Stephen Curry, Zion Williamson
Sido	Eminem, Eminem, Macklemore
The Cratez	DJ Khaled, Drake , Twenty One Pilots
Thüringen	Iowa, Midwestern United States, Tennessee, Tennessee
Till Lindemann	James Hetfield, James Hetfield, James Hetfield, Ozzy Osbourne
Tom Kaulitz	Adam Levine, Blink-182, Chris Martin, Green Day, Maroon 5
UEFA Champions League	Major League Soccer, NFC, NFL, National Football League, Ncaa
Udo Jürgens	Aretha Franklin, Billy Joel, Elton John, Michael Jackson, Rolling Stone, Tom Lehrer
Udo Lindenberg	Johnny Cash, Mick Jagger, Roger Taylor , Travis Barker
Ursula von der Leyen	Condoleezza Rice, Hillary Clinton, Mike Pence, Sarah Palin, Susan Rice
Volkswagen AG	Ford Motor Company, Ford Motor Company, Ford Motor Company, Ford Motor Company, Ford Motor Company
Walter Lübcke	Harvey Milk, John F. Kennedy, John Roll, Steve Scalise
Weimarer Republik	America, Confederation Period, Congress of the Confederation, Counterculture of the 1960s, The Confederate States of America
Westdeutscher Rundfunk Köln	ABC News, NBC, NPR

Wien	Austin Texas, Richmond Virginia, Toronto, Washington D.C.
Wilhelm II.	William Howard Taft, Woodrow Wilson, Woodrow Wilson
Wolfgang Amadeus Mozart	Alan Menken, Elvis Presley, Leonard Bernstein
ZDF	NPR, NPR, National Public Radio, PBS, PBS
Österreich	Canada, Mexico, Texas, Texas, United States
Ötzi	Spirit Cave mummy, Spirit Cave mummy, Spirit Cave mummy, Sue

Table A.3: Top Wikipedia German→American adaptations.

Entity	Human Adaptation: Wikipedia American→German
13 Reasons Why	Club der roten Bänder, Gute Zeiten schlechte Zeiten, Lammbock, Türkisch für Anfänger
Albert Einstein	Albert Einstein, Albert Einstein, Albert Einstein, Max Planck, Max Planck
Alexander Hamilton	Konrad Adenauer, Max Weber, Otto von Bismarck, Otto von Bismarck
American Civil War	Deutscher Krieg, Dreißigjähriger Krieg, German Revolution of 1918–1919, German revolutions of 1848–1849
American Horror Story	Dark, Der goldene Handschuh, Good Bye Lenin!, Tintenherz
Angelina Jolie	Barbara Schöneberger, Franka Potente, Marlene Dietrich, Romy Schneider, Veronica Maria Cäcilia Ferres
Apple Inc.	BMW, Fujitsu, SAP, Siemens
Ariana Grande	Lena Meyer-Landrut, Lena Meyer-Landrut, Lena Meyer-Landrut, Sarah Connor, Sarah Connor
Arnold Schwarzenegger	Arnold Schwarzenegger, Karl Lauterbach, Matthias Steiner, Peter Maffay, Ralf Rudolf Möller
Ashton Kutcher	Florian David Fitz, Matthias Schweighöfer, Til Schweiger, Til Schweiger
Australia	Australia, Russia, Schweiz, South Africa, Österreich
Avengers Infinity War	Das Arche Noah Prinzip, Fack ju Göhte, Fantastic Four, Who Am I
Barack Obama	Angela Merkel, Angela Merkel, Angela Merkel, Helmut Schmidt, Helmut Schmidt
Beyoncé	Helene Fischer, Sarah Connor, Veronica Ferres, Xavier Naidoo, Yvonne Catterfeld
Black Mirror	Dark, Dark, Die kommenden Tage, Krabat
Blake Lively	Josefine Preuß, Maria Furtwängler, Maria Furtwängler, Til Schweiger
Brad Pitt	Florian David Fitz, Frederick Lau, Til Schweiger, Til Schweiger, Til Schweiger
Bruce Lee	Götz Georg, Henry Maske, Julian Jacobi, Max Schmeling, no one is like Bruce Lee
Caitlyn Jenner	Kristin Otto, Magdalena Neuner, Magdalena Neuner, Niklas Kaul, Ulrike Meyfarth
California	Bavaria, Bavaria, Bayern, Bayern
Camila Cabello	Helene Fischer, Lena Meyer-Landrut, Lena Meyer-Landrut, Nadja Benaissa
Canada	Austria, Italy, Schweiz, Sweden, Österreich
Cardi B	Ace Tee, Pamela Reif, Sabrina Setlur, Sarah Connor, Schwester Ewa
Charles Manson	Andreas Baader, Issa Rammo, Papst benedikt xvi, Paul Schäfer

Charlize Theron	Baran bo Odar, Josefine Preuß, Josefine Preuß, Veronica Ferres, Veronica Maria Cäcilia Ferres
Cher	Marlene Dietrich, Nena, Nena, Nena
Chris Pratt	Elyas M'Barek, Jan Josef Liefers, Matthias Schweighöfer, Ralf Moeller, Til Schweiger
Clint Eastwood	Heinz Erhardt, Klaus Kinski, Mario Adorf, Til Schweiger, Wim Wenders
Darth Vader	Adolf Hitler, Belzebub, Hagen von Tronje, Jens Maul
Donald Glover	Elyas M'Barek, Helge Schneider, Money Boy, Stefan Raab
Drake	Bushido, Cro, Falco, Fler
Dwayne Johnson	Alexander Wolfe, Arnold Schwarzenegger, Peter Alexander, Tim Wiese, Tim Wiese
Elon Musk	Alexander Samwer, August Horch, Carl Benz, Herbert Diess, Werner von Siemens
Eminem	Bushido, Kollegah, Sido, Sido, Sido
Facebook	Das Erste, Lokalisten, Lokalisten, Schüler VZ, StudiVZ, StudiVZ
Friends	Gute Zeiten schlechte Zeiten, Gzsz, Lindenstraße, Stromberg
Game of Thrones	Babylon Berlin, Babylon Berlin, Babylon Berlin, Die unendliche Geschichte, Krabat
Google	Ecosia, Fastbot, SAP, SAP, i.d.k.
Harry Potter	Die Unendliche Geschichte, Die unendliche Geschichte, Harry Potter und ein Stein, Meggie Folchart
Heath Ledger	Christoph Waltz, Florian David Fitz, Henry Blanke, Matthias Schweighöfer, Tilman Valentin Schweiger
It	Dark, Der goldene Handschuh, Die Wolke, Pandorum
Jason Momoa	Arnold Schwarzenegger, Benno Fürmann, Christoph Waltz, Elyas M'Barek, Elyas M'Barek, Elyas M'Barek
Jeff Bezos	Alexander Samwer, Beate Heister, Martin Winterkorn, Oliver Samwer
Jeffrey Dahmer	Armin Meiwes, Fritz Haarmann, Joachim Kroll, Karl Denke, Karl Denke
Jennifer Aniston	Barbara Schöneberger, Diane Kruger, Diane Kruger, Franka Potente, Iris Berben
Jennifer Lawrence	Iris Berben, Josefine Preuß, Karoline Herfurth, Ruby O. Fee
Jennifer Lopez	Heidi Klum, Helene Fischer, Jeanette Biedermann, Mandy Capristo, Sarah Connor
John Cena	Arnold Schwarzenegger, Max Schmeling, Max Schmeling, Ralf Möller
Johnny Cash	Fantastischen vier, Helge Schneider, Peter Maffay, Peter Maffay

Johnny Depp	Christoph Maria Herbst, Christoph Waltz, Cro, Til Schweiger, Xavier Naidoo
Julia Roberts	Karoline Herfurth, Maria Furtwängler, Marlene Dietrich, Marlene Dietrich
Justin Bieber	Cro, Felix Jaehn, Lukas Rieger, McFittie, Mike Singer
Keanu Reeves	Daniel Brühl, Mario Adorf, Til Schweiger, til schweiger
Kylie Jenner	Barbara Schöneberger, Heidi Klum, Karoline Einhoff, Sarah Connor, Stefanie Giesinger
Lady Gaga	Helene Fischer, Nena, Nena, Nina Hagen, Sarah Lombardi
LeBron James	Dirk Nowitzki, Dirk Nowitzki, Dirk Nowitzki, Dirk Nowitzki, Toni Kroos
Leonardo DiCaprio	Matthias Schweighöfer, Moritz Bleibtreu, Til Schweiger, Til Schweiger, Til Schweiger
Lisa Bonet	Franka Potente, Iris Berben, Karoline Herfurth, Maria Furtwängler
Madonna	Blümchen, Helene Fischer, Helene Fischer, Helene Fischer, Sarah Connor
Mark Wahlberg	Florian David Fitz, Til Schweiger, Tilman Valentin Schweiger, Alexei Alexejewitsch
Martin Luther King Jr.	Hans Scholl, Hans Scholl, Helmut Palmer, Robert Blum, Sophie Scholl
Marvel Cinematic Universe	Bavaria Film, Havelstudios, Phantásien, Rat Pack Filmproduktion, Tatort
Michael Jackson	Herbert Grönemeyer, Nena, Udo Jürgens, Xavier Naidoo, Xavier Naidoo
Mila Kunis	Josefine Preuß, Matthias Schweighöfer, Vanessa Mai
Miley Cyrus	Lena Meyer-Landrut, Lukas Rieger, Nena, Sarah Connor, Yvonne Catterfeld
Muhammad Ali	Alexander Abraham, Boris Becker, Max Schmeling, Max Schmeling, Sven Ottke
Natalie Portman	Barbara Schöneberger, Diane Kruger, Franka Potente, Iris Berben
New York City	Berlin, Berlin, Berlin, Berlin, Frankfurt
Nicole Kidman	Evelyn Hamann, Franka Potente, Senta Berger, iris berben
Peaky Blinders	Dark, Dieter Schwarz, Im Westen Nichts Neues, Tatort, Tatort
Philippines	Greece, Griechenland, Mallorca, Mallorca
Post Malone	Bushido, Bushido, Cro, Cro, Kollegah
Rihanna	Helene Fischer, Lena Meyer-Landrut, Lena Meyer-Landrut, Nena
Riverdale	Babylon Berlin, Berlin Tag und Nacht, Neues vom Süderhof, Türkisch für Anfänger
Robert Downey Jr.	Christoph Waltz, Günter Strack, Martin Semmelrogge, Moritz Bleibtreu, Til Schweiger

Robin Williams	Hape Kerkeling, Heinz Erhardt, Peter Maffay, Silvia Seidel, Tim Bendzko
Ronald Reagan	Helmut Schmidt, Konrad Adenauer, Konrad Adenauer, Konrad Adenauer
Ryan Reynolds	Daniel Brühl, Florian David Fitz, Matthias Schweighöfer, Til Schweiger, Til Schweiger
Scarlett Johansson	Lena Gercke, Romy Schneider, Sarah Connor, Sarah Connor, Veronica Ferres
Selena Gomez	Lena Meyer-Landrut, Lena Meyer-Landrut, Nena, Nora Tschirner
September 11 attacks	Anschlag im OEZ, Dresden Bombing, Mauerfall, RAF-Attentate, Terroranschlag Olympia 1972
Shaquille O'Neal	Dirk Nowitzki, Dirk Nowitzki, Mehmet Scholl, Niklas Süle
Star Wars	Dark, Metropolis, Traumschiff Surprise – Periode 1, Who Am I?, i.d.k
Stephen Curry	Dirk Nowitzki, Dirk Nowitzki, Dirk Nowitzki, Dirk Nowitzki, Manuel Neuer
Stranger Things	8 Tage, Babylon Berlin, Dark, Tatort, Tatort
Sylvester Stallone	Henry Blanke, Jan Josef Liefers, Michael Bully Herbig, Michael Fassbender, Til Schweiger
Taylor Swift	Lena Meyer-Landrut, Lena Meyer-Landrut, Sarah Connor, Sarah Connor, Yvonne Catterfeld
Ted Bundy	Joachim Kroll, Josef Fritzl, Niels Högel, Rudolf Pleil, Rudolf Pleil
The Big Bang Theory	Doctor's Diary, Stromberg, Stromberg, der Tatortreiniger
The Crown	Babylon Berlin, Deutschland 83, Die Deutschen, Karl der Große
The Handmaid's Tale	Dark, Dark, Der Pass, Die Wanderhure, Er ist wieder da
The Walking Dead	Dark, Dark, Der goldene Handschuh, Zombies From Outer Space
Tom Brady	Franz Beckenbauer, Michael Ballack, Oliver Kahn, Thomas Müller, Uli Stein
Tom Cruise	Benno Fürmann, Benno Fürmann, Christoph Waltz, Elyas M'Barek, Matthias Schweighöfer
Tom Hanks	Christoph Waltz, Christoph Waltz, Daniel Brühl, Til Schweiger
Tom Hardy	Bruno Ganz, Michael Herbig, Til Schweiger, Wotan Wilke Möhring
Tom Holland	Daniel Brühl, Frederick Lau, Matthias Schweighöfer, Matthias Schweighöfer, Til Schweiger
Tupac Shakur	Farid Bang, Haftbefehl, Kollegah, Kristoffer Klauß, Peter Fox
United States	BRD, Bundesrepublik Deutschland, Deutschland, Germany, Germany
Vietnam War	Berlin Wall, First world war, Kosovokrieg, World War II

Wikipedia	Brockhaus, Brockhaus Enzyklopädie, Brockhaus Enzyklopädie, Duden, dict.cc
Will Smith	Daniel Brühl, Elyas M'Barek, Sascha Reimann, Sido, Til Schweiger
X-Men	Abwärts, Fantastic Four, Freaks, Krabat, Who Am I
YouTube	Lokalisten, MyVideo, MyVideo, ProSieben, lokalisten
Zac Efron	Frederick Lau, Lukas Rieger, Peter Kraus, Walter Sedlmayr
Zendaya	Franka Potente, Iris Berben, Lena Meyer-Landrut, Lena Meyer-Landrut, Yvonne Catterfeld

Table A.4: Top Wikipedia American→German adaptations.

Entity	Top Five WikiData Adaptations
Abraham Lincoln	Victor Adler, Johann Joachim Christoph Bode, Willem Barentsz, Hermann Wagener, Robert von Mohl
Al Capone	Hans H. Zerlett, Fritz Thyssen, Adam Rainer, Franz Winkelmeier, Christian Louis, Duke of Brunswick-Lüneburg
Alfred Hitchcock	Edgar Reitz, Jan Josef Liefers, Mario Adorf, Max Frisch, Armin Mueller-Stahl
Benedict Arnold	Hans-Georg Hess, Isabelle Eberhardt, Günther Heydemann, Max Schreck, Louis Blenker
Bill Gates	Ferdinand von Zeppelin, Günther Jauch, Nikolaus Harnoncourt, Sepp Blatter, Alfred Grosser
Britney Spears	Herta Müller, Günter Grass, Joachim Gauck, Hans-Dietrich Genscher, Koča Popović
Donald Trump	Max Frisch, Thomas Gottschalk, Jan Josef Liefers, Rainer Werner Fassbinder, Christa Wolf
Elvis Presley	Reinhard Lakomy, James Last, Herbert Achternbusch, Fritz Hauser, Hans-Peter Pfammatter
Ernest Hemingway	Karlheinz Böhm, Ricarda Huch, Michael Ballhaus, Arnold Zweig, Michael Fassbender
Frank Lloyd Wright	Ferdinand Hodler, Johan Zoffany, Hans Thoma, Arne Jacobsen, Lucas Cranach the Younger
George Washington	Friedrich Wilhelm von Seydlitz, Dagobert Sigmund von Wurmser, Heinz Guderian, Ernst Gideon von Laudon, George Olivier, count of Wallis
Henry Ford	Heinz Sielmann, Wieland Schmied, Manfred Krug, Paul Maar, Armin Mueller-Stahl
Hillary Clinton	Pope Benedict XVI, Willy Brandt, Angela Merkel, Helmut Schmidt, Kurt Biedenkopf
Homer Simpson	Elizabeth Lavenza, Hans Fugger, Baron Strucker, Herbert of Wetterau, Prince Johannes of Liechtenstein
Jimi Hendrix	Marius Müller-Westernhagen, Karl Richter, Reinhard Lakomy, Michael Cretu, Paul van Dyk
Kim Kardashian	Erika Mann, Frank Wedekind, Til Schweiger, Fritz von Opel, Carmen Electra
Marilyn Monroe	Gerhart M. Riegner, Viktor de Kowa, Otto Sander, Hans Hass, Dorothee Sölle
Michael Jordan	Jean-Claude Juncker, Richard von Weizsäcker, Herta Müller, Konrad Adenauer, Helmut Kohl
Louis Armstrong	Herbert Prikopa, Till Lindemann, Nico, Klaus Voormann, Jakob Adlung
Neil Armstrong	Stefan Hell, Franz-Ulrich Hartl, Reinhard Genzel, Charles Weissmann, Harald zur Hausen
Noam Chomsky	Günter Grass, Herta Müller, Heinrich Böll, Peter Handke, Juli Zeh

Oprah Winfrey	Günter Grass, Peter Scholl-Latour, Elfriede Jelinek, Juli Zeh, Christa Wolf
Orville Wright	Frank Thiess, Jessica Hausner, Elmar Wepper, Wolf Jobst Siedler, Marc Rothemund
Richard Nixon	Heinrich von Brentano, Ernst Benda, Gustav Heinemann, Heiner Geißler, Heinrich Albertz
Superman	Magneto, Nightcrawler, Sinterklaas, Silent Night, Victor Frankenstein
Steve Jobs	Victor Klemperer, Joschka Fischer, Jürgen Kuczynski, Joachim Fest, Dieter Hallervorden
Steven Spielberg	Herta Müller, Jean-Claude Juncker, Hans-Dietrich Genscher, Joachim Gauck, Koča Popović
Tiger Woods	Charles Dutoit, Shania Twain, Lise Meitner, Michael Haneke, Otto Hahn
Walt Disney	Shania Twain, Charles Dutoit, Lise Meitner, Otto Hahn, Michael Haneke
John F. Kennedy	Bernhard von Bülow, Otto von Habsburg, Hans-Jochen Vogel, Prince Henry of Prussia, Frederick Augustus III of Saxony
Charles Lindbergh	Pina Bausch, Ferdinand von Zeppelin, Nikolaus Harnoncourt, Jan Josef Liefers, Wolf Biermann
Rosa Parks	Hermann Lenz, Wilhelm Feldberg, Horst Tappert, Peter Stein, Gert Jonke
Serena Williams	Charles Dutoit, Lise Meitner, Michael Haneke, Richard von Coudenhove-Kalergi, Klaus Clusius

Table A.5: We show top-5 predictions out of the top-100 for American→German adaptations on the Veale NOC subset using **WikiData**. These are compared to our human annotations in our results.

Entity	Top Five 3CosAdd Adaptations: American→German adaptations on the Veale NOC
Abraham Lincoln	Napoleon, Napoléon Bonaparte, Erzherzog Johann, Otto von Bismarck, Kaiser Wilhelm II.
Al Capone	Nazis, SA-Mann, Verhaftungswellen, Judenverfolgung, Fluchthilfe
Alfred Hitchcock	Fritz Lang, Helmut Käutner, Willi Forst, Emil Jannings, Heinz Rühmann
Benedict Arnold	Russlandfeldzug 1812, Schlacht bei Roßbach, Jean-Victor Moreau, schwedischen Armee, Alexander Wassiljewitsch Suworow
Bill Gates	congstar, Alnatura, GMX, ChessBase, Gardeur
Britney Spears	Glasperlenspiel, Unheilig, Helene Fischer, Christina Aguilera, Herbert Grönemeyer
Charles Lindbergh	Segelflieger, Flugpioniere, Zeppelins, Adolf Hitler, Caproni
Donald Trump	Deutschland, Österreich, Trump, Strache, Bundestagswahlkampf
Elvis Presley	Udo Jürgens, Elvis Presley, Hits, den Beatles, der Beatles
Ernest Hemingway	Stefan Zweig, Franz Werfel, Joachim Ringelnatz, Hermann Hesse, Gottfried Benn
Frank Lloyd Wright	Adolf Loos, Le Corbusier, Bruno Schmitz, Entwürfen, Fritz Höger
George Washington	Napoléon Bonaparte, Friedrich dem Großen, Napoleon, Friedrich der Große, Napoleon Bonaparte
Henry Ford	Ferdinand Porsche, Büssing, Krupp, Ettore Bugatti, Steyr-Daimler-Puch
Hillary Clinton	Deutschland, Bundestagswahlkampf, Österreich, Sarkozy, Strache
Homer Simpson	Eingangsszene, verulkt, Schlusssequenz, Off-Stimme, Muminfamilie
Jack The Ripper:Ripper	Tat, Werwolf, Täter, Dritten Reich, Mörder
Jay Z	Xavier Naidoo, D-Bo, Sido, Rosenstolz, David Guetta
Jimi Hendrix	Udo Jürgens, Tangerine Dream, Jimi Hendrix, Pink Floyd, Depeche Mode
John F. Kennedy	Adolf Hitler, Bundeskanzlers, Adolf Hitlers, Adolf Hitler, Hitler
Kim Kardashian	Kaas, gotv, Frank Zander, Herbert Grönemeyer, Roland Kaiser
Louis Armstrong	Richard Tauber, Django Reinhardt, Udo Jürgens, Sidney Bechet, Jazzorchester
Marilyn Monroe	Marlene Dietrich, Lil Dagover, Elisabeth Bergner, Brigitte Bardot, Romy Schneider
Michael Jordan	Powerplay, Xavi, Predrag Mijatović, NHL-Historie, Franck Ribéry

Neil Armstrong	Juri Gagarin, Vorbeiflug, Weltraum, Raumstation Mir, Raumfahrer
Noam Chomsky	Jürgen Habermas, Hans-Ulrich Wehler, Carl Schmitt, Theodor W. Adorno, Norbert Elias
Oprah Winfrey	Harald Schmidt, Thomas Gottschalk, Satiresendung, ORF-Sendung, Hape Kerkeling
Orville Wright	Parseval, Luft Hansa, Hugo Junkers, Ernst Heinkel, Claude Dornier
Richard Nixon	Österreich, Deutschland, Bundeskanzler, Bundeskanzlers, Bundespräsidenten
Rosa Parks	NS-Militärjustiz, Franz Jägerstätter, NS-Opfer, Bücherverbrennung, Baum-Gruppe
Serena Williams	Dick Jaspers, Philipp Kohlschreiber, Semifinale, Achtelfinale, Dominic Thiem
Steve Jobs	Steve Jobs, Sony, Electronic Arts, Netscape, Atari
Steven Spielberg	Hörspielproduktion, Helmut Käutner, Fellini, Oliver Hirschbiegel, Kinofilm
Superman	Superman, Batman, Superhelden, Monster, Spider-Man
Tiger Woods	Rekordeuropameister, Österreich, spanische Team, ÖFB-Cupsieger, Deutschland
Walt Disney	Fritz Lang, Sascha-Film, Fellini, UFA, "Das Cabinet des Dr. Caligari"

Table A.6: We show top-5 predictions out of the top-100 for American→German adaptations on the Veale NOC subset using **3CosAdd**. These are compared to our human annotations in our results.

Entity	Top Five Learned Adaptations: American→German adaptations on the Veale NOC
Abraham Lincoln	Konrad Adenauer, Helmut Schmidt, Willy Brandt, Helmut Kohl, Adenauer
Al Capone	Andreas Baader, Leo Katzenberger, Paul Schäfer, Strippel, Hermann Langbein
Alfred Hitchcock	Helmut Käutner, Til Schweiger, Mario Adorf, Paul Verhoeven, Dennis Hopper
Benedict Arnold	Otto von Bismarck, Bismarcks, Bismarck, Preußens, Kaiserreiches
Bill Gates	Martin Winterkorn, Volkswagen AG, Daimler-Chrysler, Robert Bosch GmbH, Volkswagen AG
Britney Spears	Sarah Connor, Nena, Helene Fischer, Lena Meyer-Landrut, Moses Pelham
Charles Lindbergh	Chaim Weizmann, Tomáš Garrigue Masaryk, Ferdinand Sauerbruch, Fritz Haber, Chaim Arlosoroff
Donald Trump	Helmut Schmidt, Angela Merkel, Gerhard Schröder, Helmut Kohl, Bundesaußenminister
Elvis Presley	Udo Jürgens, Peter Maffay, Cliff Richard, Achim Reichel, Lou Reed
Ernest Hemingway	Paul Schlenther, Marcel Reich-Ranicki., Timothy Leary, Erwin Leiser, Alice Walker
Frank Lloyd Wright	Albert Einstein, Max Planck, Max Born, Hermann von Helmholtz, Arnold Sommerfeld
George Washington	Otto von Bismarck, Otto von Bismarck, Konrad Adenauer, Engelbert Dollfuß, Joseph Wirth
Henry Ford	Ernst Abbe, Carl Duisberg, Bubbe, Aby Warburg, Sybel
Hillary Clinton	Angela Merkel, Angela Merkel, Helmut Schmidt, Gerhard Schröder, Bundesinnenminister
Homer Simpson	Rolf Hochhuth, Carl Bernstein, Uwe Tellkamp, Wolfgang Völz, Richard Gere
Jack The Ripper:Ripper	Sarah Connor, Spike Jonze, Timberlake, "Das Urteil", "Nichts als die Wahrheit"
Jay Z	will.i.am, Moses Pelham, Silbermond, Xavier Naidoo, Kanye West
Jimi Hendrix	Peter Maffay, Udo Lindenberg, Depeche Mode, Xavier Naidoo, Die Toten Hosen
John F. Kennedy	Konrad Adenauer, Helmut Schmidt, Willy Brandt, Helmut Kohl, Bundeskanzler
Kim Kardashian	Heidi Klum, Ruth Moschner, Ellen DeGeneres, Circus HalliGalli, Oliver Pocher
Louis Armstrong	Peter Maffay, Radioaufnahmen, Udo Lindenberg, Achim Reichel, Helge Schneider
Marilyn Monroe	Walter Giller, Jessica Tandy, Liv Ullmann, Edgar Selge, Betty White
Michael Jordan	Dirk Nowitzki, Toni Kroos, Zlatan Ibrahimović, Xavi, Zinédine Zidane

Neil Armstrong	Max von Laue, Albert Einstein, Chaim Weizmann, Johannes R. Becher, Ernst Abbe
Noam Chomsky	Albert Einstein, Nobelpreisträger, Max Planck, American Psychological Association, Hans Bethe
Oprah Winfrey	Anja Kling, "Forsthaus Falkenau", Uschi Glas, "Saturday Night Live"., Anke Engelke
Orville Wright	Kawaishi, Rjabuschinski, Monistenbund, Dethmann, Leo Baeck Instituts
Richard Nixon	Helmut Schmidt, Konrad Adenauer, Willy Brandt, Helmut Kohl, Gerhard Schröder
Rosa Parks	Sophie Scholl, Die letzten Tage, Emil Jannings., Ruth Wilson, Monica Bleibtreu
Serena Williams	Max Schmeling, Wilfried Dietrich, Gottfried von Cramm, Henry Maske, László Kubala
Steve Jobs	DaimlerChrysler, Volkswagen, Siemens, Sanyo, Fujitsu
Steven Spielberg	Til Schweiger, Ethan Hawke, Matthias Schweighöfer, Samuel L. Jackson, Ryan Reynolds
Superman	Jabberwocky, Freaks, Scarface, Leatherface, Krabat
Tiger Woods	Dirk Nowitzki, deutschen U21-Nationalmannschaft, MTV Gießen, Mats Hummels, Franz Beckenbauer
Walt Disney	Helmut Dietl, Peter Ustinov, David Mamet, Rainer Werner Fassbinder, Sönke Wortmann

Table A.7: We show top-5 predictions out of the top-100 for American→German adaptations on the Veale NOC subset with our **Learned Adaptation** approach. These are compared to our human annotations in our results.

Appendix B: Diplomacy

Our Diplomacy (Chapter 7) appendix contains:

1. examples of game summaries written by players (Table B.1);
2. the game engine view of the board (Figure B.1);
3. examples of persuasion techniques (Table B.2);
4. Harbingers word lists that are used as features in the logistic regression model (Table B.3); and
5. A full transcript between two players, Germany and Italy (Table B.4). Messages are long and carefully composed. This transcript is from the game described in Section 7.2.1 (Warning: it is dozens of pages).

B.1 Further Details

User	Summary
Italy	This was an interesting game, with some quality play all around, but I felt like I was playing harder than most of the others. I felt early on that I could count on Austria remaining loyal, which worked to my benefit, as it allowed me freedom to stab and defeat a very strong French player before he got his legs under him. At the same time, Austria was a little too generous in granting me centers and inviting me to come help him against Russia, which allowed me to take advantage once I was established in the Middle Atlantic.
Russia	Definitely a good game by Italy - which is interesting to me, because his initial press struck me as erratic and aggressive, making me not want to work with him. I'm curious if the same negotiating approach was taken with the other players who did work with him early on, or if he used a different negotiating approach with closer neighbors.

Table B.1: Users optionally provide free response descriptions of the game. This can be used for qualitative analysis or potentially for algorithmic summarization.

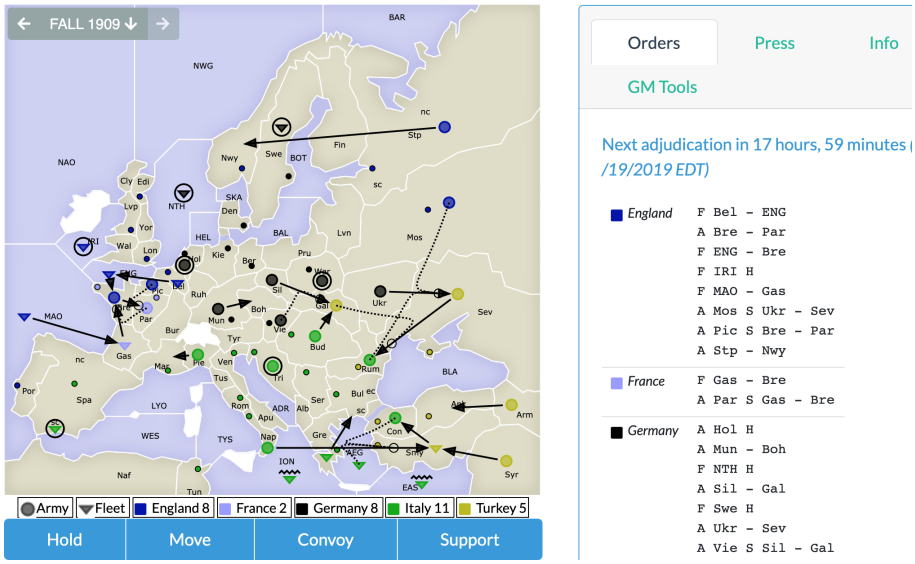

Figure B.1: The board game as implemented by Backstabbr. Players place moves on the board and the interface is scraped.

Principle	Example
Authority	Sent to Germany, England, Austria, Russia: So, England, Germany, Russia, y'all played a great turn last turn. You got me to stab my long-time ally and you ended our pretty excellent 7-year run as an alliance. Russia told me he was with me if I stab Austria. England told me he wanted me to solo so long as I would "teach him" and help his along to second place. Then y'all pulled the rug out from under me. It was clever and effective. At this stage, my excitement about the game has diminished quite a bit. And of course I'm happy to play on and take my lumps for falling for "Hey, I really want you to solo, just help me place second," but if you guys just want to call it a five-way draw among us and grab a beer together, while reviewing the statistics, that's really my preference. I am outnumbered and I obviously can't solo. And I'm sure some of you in the north are eager to send everyone else flying my way, but I expect Russia and England to be careful, and so I'm not sure there is much room to move forward without simply tipping the board to Germany's favor. I propose that we draw and hug it out.
Reciprocity	1) You've been straight with me all game. 2) You have a much better ability to read the board than she does. 3) You're on the other side, so you can't really stab me, but I could totally see her moving to Tyrolia some time soon. 4) You're not in France's pocket.
Likability	Maine is beautiful! I used to go to scout camp there.
Scarcity	I'd like to have your final thoughts on A/R as quickly as possible so that I have time to execute a plan. But I understand if you want time to think about it.

Table B.2: Examples of persuasion from the games annotated with tactics from Cialdini and Goldstein (2004).

Feature	Key Word
claim	accordingly, as a result, consequently, conclude that, clearly, demonstrates that, entails, follows that, hence, however, implies, in fact, in my opinion, in short, in conclusion, indicates that, it follows that, it is highly probable that, it is my contention, it should be clear that, I believe, I mean, I think, must be that, on the contrary, points to the conclusions, proves that, shows that, so, suggests that, the most obvious explanation', "the point I'm trying to make", 'therefore, thus, the truth of the matter, to sum up, we may deduce
subjectivity	abandoned, abandonment, abandon, abase, abasement, abash, abate, abdicate, aberration, aberration, abhor, abhor, abhorred, abhorrence, abhorrent, abhorrently, abhors, abhors, abidance, abidance, abide, abject, abjectly, abjure, abilities, ability, able, abnormal, abolish, abominable, abominably, abominate, abomination, above, above-average, abound, abrade, abrasive, abrupt, abscond, absence, absentee, absent-minded, absolve, absolute, absolutely, absorbed, absurd, absurdity, absurdly, absurdness, abundant, abundance, abuse, abuse, abuse, abuses, abuses, abusive, abysmal, abysmally, abyss, accede, accentuate, accept, acceptance, acceptable, accessible, accidental, acclaim, acclaim, acclaimed, acclamation, accolade, accolades, accommodative, accomplish, accomplishment, accomplishments, accord, accordance, accordantly, accost, accountable, accurate, accurately, accursed, accusation, accusation, accusations, accusations, accuse, accuses, accusing, accusingly, acerbate, acerbic, acerbically, ache, achievable, achieve, achievement, achievements, acknowledge, acknowledgement, acquit, acrid, acridly, acridness, acrimonious, acrimoniously, acrimony, active, activist, activist, actual, actuality, actually, acumen, adamant, adamantly, adaptable, adaptability, adaptive, addict, addiction, adept, adeptly, adequate, adherence, adherent, adhesion, admirable, admirer, admirable, admirably, admiration, admire, admiring, admiringly, admission, admission, admit, admittedly, admonish, admonisher, admonishingly, admonishment, admonition' ...
expansion	additionally, also, alternatively, although, as an alternative, as if, as though, as well, besides, either or, else, except, finally, for example, for instance, further, furthermore, however, in addition, in fact, in other words, in particular, in short, in sum, in the end, in turn, indeed, instead, later, lest, likewise, meantime, meanwhile, moreover, much as, neither nor, next, nonetheless, nor, on the other hand, otherwise, overall, plus, rather, separately, similarly, specifically, then, ultimately, unless, until, when, while, yet
contingency	accordingly, as a result, as long as, because, consequently, hence, if and when, if then, in the end, in turn, indeed, insofar as, lest, now that, once, since, so that, then, thereby, therefore, thus, unless, until, when
premise	after all, assuming that, as, as indicated by, as shown, besides, because, deduced, derived from, due to, firstly, follows from, for, for example, for instance, for one thing, for the reason that, furthermore, given that, in addition, in light of, in that, in view of, in view of the fact that, indicated by, is supported by, may be inferred, moreover, owing to, researchers found that, secondly, this can be seen from since, since the evidence is, what's more, whereas

temporal-future	after, afterward, as soon as, by then, finally, in the end, later, next, once, then, thereafter, till, ultimately, until
temporal-other	also, as long as, before, before and after, earlier, in turn, meantime, meanwhile, now that, previously, simultaneously, since, still, when, when and if, while
comparisons	after, although, as if, as though, besides, by comparison, by contrast, conversely, earlier, however, in contrast, in fact, in the end, indeed, instead, meanwhile, much as, neither nor, nevertheless, nonetheless, nor, on the contrary, on the one hand on the other hand, on the other hand, previously, rather, regardless, still, then, though, when, whereas, while, yet

Table B.3: The word lists used for our Harbingers (Niculae et al., 2015) logistic regression models.

B.2 A Full Game Example

#	Speaker	Message	Actual Lie	Suspected Lie
0	Italy	Germany! Just the person I want to speak with. I have a somewhat crazy idea that I've always wanted to try with I/G, but I've never actually convinced the other guy to try it. And, what's worse, it might make you suspicious of me. So...do I suggest it? I'm thinking that this is a low stakes game, not a tournament or anything, and an interesting and unusual move set might make it more fun? That's my hope anyway. What is your appetite like for unusual and crazy?	Truth	Truth
1	Germany	You've whet my appetite, Italy. What's the suggestion?	Truth	Truth
2	Italy	Okay, don't hate me! Key West (Just thought of the name lol) Basic point is that I move to Tyr in Spring and into Mun in the Fall, while I take Tun with my fleet. I build A Ven/F Nap. You open to Ruh/Hol/Kie, and force Belgium. You wind up with 2 builds, and the sympathy and concern of your neighbors who are astonished at the crazy Italian. "What a stupid move, he can't hold Munich!" Trap is set to obliterate France in the Spring of 02. Bel S Mun - Bur, Ven - Pie, Tun - WMed. France won't see it coming, He will see that attack on Munich and think that both you and I will be occupied for a while. So Spring 02 should be a serious surprise. Now, you're taking risk here, because you're giving up a home center for a turn hence the "Key"), but I think you can see pretty clearly that I derive no benefit from trying to double-cross you. After all, Italy trying to hold Munich is just dumb. I'm from a school of thought that even trying to move to Munich is just dumb. But this would be the one exception. I can't hold Munich, and even if I wanted to, it would give me an awkward snake formation in the middle of the board that is a great way to be first eliminated. So I think this works because you know (even more than Austria in a traditional Key Lepanto), that I'm not going to stab you. And doing it this way allows us to take Burgundy by surprise, it ensures you get Belgium, and it crushes your biggest rival and toughest border (France). Plus, it'll be fun. The Key West! Thoughts? **This message crashed the Beta version of the bot due to its length and is not in the dataset. It is manually extracted and added here the purposes of readability**	Truth	None

3	Germany	It seems like there are a lot of ways that could go wrong...I don't see why France would see you approaching/taking Munich–while I do nothing about it–and not immediately feel skittish	Truth	Truth
4	Italy	Yeah, I can't say I've tried it and it works, cause I've never tried it or seen it. But how I think it would work is (a) my Spring move looks like an attack on Austria, so it would not be surprising if you did not cover Munich. Then (b) you build two armies, which looks like we're really at war and you're going to eject me. Then we launch the attack in Spring. So there is really no part of this that would raise alarm bells with France. All that said, I've literally never done it before, and it does involve risk for you, so I'm not offended or concerned if it's just not for you. I'm happy to play more conventionally too. Up to you.	Truth	None
5	Italy	I am just sensing that you don't like this idea, so shall we talk about something else? That was just a crazy idea I've always wanted to try. I'm happy to play more conservatively.	Truth	None
6	Italy	Any thoughts?	Truth	None
7	Germany	Sorry Italy I've been away doing, um, German things. Brewing Lagers?	Truth	Truth
8	Germany	I don't think I'm ready to go for that idea, however I'd be down for some good ol'-fashioned Austria-kicking?	Truth	Lie
9	Italy	I am pretty conflicted about whether to guess that you were telling the truth or lying about the "brewing lagers" thing. I am going to take it literally and say thumbs down even though I don't think you meant it deceptively.	Truth	Truth
10	Italy	But I think I can get over "Lagergate" and we can still be friends. As of right now, I think Austria may be my most reliable ally. I'm thinking I'd like to play as a Central Trio if you have any interest in that. Thoughts?	Truth	Truth
11	Germany	We haven't even passed a season yet and you have a 'most reliable ally'? I'll consider this proposal but, basically, I'm not going to expose myself to risk from either of you until I've seen a bit of your behavior	Truth	Truth

12	Italy	Well, at least I have an idea of who to trust. Obviously, my ideas are subject to change. I understand your desire to watch behavior before committing to anything. I, personally, am a partner player. I look carefully early in the game for a small group to work with, and then I value loyalty and collaboration. I like to work closely with a tight-knit alliance. If you prefer to hop and back and forth, or play more of an individual game, then we might not be a good match. I'm looking for a loyal ally or two that I can coordinate with and make awesome moves with. Makes the game easier and a lot more fun.	Truth	Truth
13	Italy	Just an FYI: I've now had both England and France suggest to me that I should move to Tyrolia and France will support me to Munich in the Fall. One saying that to me is not a big deal, but with both mentioning it, my alarm bells are going off. I am concerned about an E/F. I'm certainly not moving to Tyrolia. But I just want you to be cautious here. I feel like England and France are working together.	Truth	Truth
14	Germany	I appreciate the tip, but I'm wondering why you're so against ousting me from Munich if I haven't explicitly agreed to be your ally?	Truth	Truth
15	Italy	Because it is terrible, terrible play for Italy to attack Germany, in my view. If I were to attack you in Munich, I could never hold Munich. So, all I would be doing is weakening you, and helping France, England, or both to get really big. I don't have any long-term path going north. Helping France to take you down is a sucker's play, whether you are working with me or not.	Truth	None
16	Italy	Did France tell you he was moving to Burgundy, or was that a stab?	Truth	Truth
17	Germany	I was not informed of it, no. And England is leading me to believe it's part of a play for Belgium, so if they're working together this might be a trick... Italy, you seem like a straight shooter, and Austria has confirmed with me about your two's alliance. So I'll confide in you–this is my first ever game of diplomacy, and I think that teaming up with the two of you could help me learn more and have more fun. So, if you're still interested in a central powers alliance, I'm in.	Truth	Truth
18	Germany	Okay full disclosure: I'm not very smart, and I accidentally let slip to England that you told me France was plotting to take Munich. I'm sorry for the error but I figured it was better to admit it so you know that England/France may not trust you.	Truth	Truth
19	Italy	Okay, thanks for telling me.	Truth	Truth
20	Germany	So, um, no alliance then?	Truth	Truth

21	Italy	I do want to be allies. Sorry, busy weekend here running around with bambinos. More to come.	Truth	Truth
22	Germany	What would you think of helping me take Marseilles in two turns?	Truth	Truth
23	Italy	Hi Germany, I'll certainly consider that. Though, I'll note: traditionally, Germany would help Italy to Marseilles if the two of them work together there. The reason is that: if I help you to Marseilles, I'm basically cut off from going west and getting anything myself. So, usually, Germany would help Italy into Marseilles to encourage Italy to come west and Germany would plan to take Paris, Belgium and Brest.	Truth	Truth
24	Germany	Fair enough–I'll help you take it, then, but I'll need to deal with Belgium first.	Truth	Truth
25	Italy	How are things going with England? I think that getting him to work with you is the main key here.	Truth	Truth
26	Germany	I'm trying–I just offered to assist with taking Sweden in exchange for some assistance into Belgium...not sure if they'll go for it...	Truth	Truth
27	Italy	I'll check with England and try to see where his head is at.	Truth	Truth
28	Germany	I've actually been thinking about this game all day and have come up with a plan I like a bit better... but England still hasn't responded to my initial offer.	Truth	None
29	Italy	That's the worst! And I'm glad to see you're so focused on this in your first game. It's a really great game if you put in the time and effort!	Truth	Truth
30	Germany	You're definitely telling the truth on that one. So can I count on you to move to piedmont this season?	Truth	Truth
31	Italy	I don't think I can afford to move to Piedmont this season. I don't really trust Austria to avoid walking through that door if I leave it wide open. I think you need to get England on board to attack France.	Truth	Truth
32	Germany	That's valid. And actually I was conferring with England and we concluded that it's not really gonna be possible for me to help you take Marseilles this year anyway. ...what are you and Austria planning for this year, then? I'm willing to tell you my plans in exchange as a gesture of trust. Have you communicated at all with England or France?	Truth	Truth
33	Italy	Hi, are you there? Just woke up. England did return my message, but he did not tell me anything substantive so I really don't know what he's doing. I'm planning to move towards Turkey.	Truth	Truth

34	Italy	Well, you're in trouble. That England move is trouble. I'm going to try to convince him to change course. I suggest you be very kind to him, and don't burn that bridge. I think your game hinges on turning England around.	Truth	Truth
35	Italy	Hi Germany, I'm working hard on turning England. And I'm also trying to get Russia to come to your aid. Doing the best I can! I'll keep you posted.	Truth	Truth
36	Germany	England just told me that Russia is helping them to take Denmark so that may be a lost cause. Granted, the source for that intel is a serpentine jackal-spawn	Truth	Truth
37	Italy	Okay, I'm reasonably sure that England wants to take the Channel and attack France now. I believe that you should basically do whatever England asks to help make this happen. As long as E attacks F, you will be in a much better position, and you'll gain back centers quickly. What are you hearing?	Truth	Truth
38	Germany	What are your plans for this turn? I can't help but notice that Munich is surrounded by foreign armies on three sides... I wish I could be more helpful but I'm pretty much just treading water right now trying not to lose anything else	Truth	Truth
39	Italy	Hey — sorry, just getting back into this now.	Truth	Truth
40	Italy	I have good news! (1) I am finally attacking France this turn. (2) I will be supporting Munich to hold from Tyrolia. Let's turn this game around, yes?	Truth	Truth
41	Italy	I am pretty sure that England is not attacking you this turn. And I am committed to supporting Munich holding. Make sure you don't move Munich so that it can take my support.	Truth	Truth
42	Germany	Okay, can do. Thanks!	Truth	Truth
43	Italy	I suggest that you order: Kiel Support Berlin holding Berlin Support Munich holding Helg to Holland Munich Support Berlin holding	Truth	Truth
44	Germany	I agree completely–although I didn't know that a country could hold *and* support at the same time! Thanks!	Truth	Truth
45	Germany	Thanks Italy. Hope you're enjoying the weather on the Anatolian	Truth	Truth
46	Italy	I will be supporting Munich to hold again. And I'll be trying to get Russia to back off your flank and protect himself against an Austrian stab that is coming.	Truth	None

47	Italy	Two bits of advice: #1 I suggest you tell Russia that Austria is coming for him. You really want Russia to move Sil back to Gal. You might also suggest to Russia that is he supports you to Denmark, you will then support Russia back to Sweden. I don't know yet if it actually makes sense to do that, but you want Russia thinking that you are eager to work with him. He'll be hoping for a reason to break off his attack on you at this point.	Truth	None
48	Italy	#2 Here is the move set I would suggest right now: Kiel Support Holland holding Holland Support Wales to Belgium (tell England you are going to order this support and he can take it or leave it) Munich Support Berlin holding Berlin Support Munich holding I think that both France and Russia are about to back off you, as they are both under fire at home. Just hold still, and soon you should be able to break out of this holding pattern.	Truth	None
49	Germany	God, I hope so! I'm attempting to make that deal with russia now...and I'm talking with England re: Belgium	Truth	Truth
50	Italy	It's none of my business, but if you do plan to take Denmark, I strongly recommend you wait until Fall. I think the most important thing for you right now is getting England fully committed against France. If that happens, taking Denmark later will be easy.	Truth	Truth
51	Germany	I think me and England are really on the same page at this point regarding France. I'm actually sort of running counter-intelligence for England (and my friends to the south, of course!) with Russia right now. England and I talked about Denmark too...and it seems like one or the other of Denmark or Belgium should work out for me this year and I'm fine with that	Truth	Truth
52	Italy	Great to hear. Thank you.	Truth	Truth
53	Germany	Do you need me to disrupt Bur this year? I'll need to seriously trust Russia if I'm going to risk not holding my eastern front, I think...	Truth	Truth
54	Italy	I do think a move to Burgundy makes sense for you this turn, and I can't imagine Russia attacking you here. He has a serious Austria problem. I suggest this: Mun - Bur Ruh - Bel Hol Support Ruh - Bel Ber - Kie Tell Russia that the last thing in the world you want to see is Austria run him over, and you're willing to help keep Russia viable if necessary (you're angling for Russia to disband his northern holdings this turn).	Truth	Truth
55	Italy	And ask England nicely to support Ruh - Hol, with the explanation that you don't plan to ask for Denmark back, but you think it would help you both to diminish France. (You'll get Den back eventually, but you want England to think you don't care about it).	Truth	Truth

56	Germany	Thanks, I'll work on these. ...Why didn't you scooch into the Aegean behind Austria? You could have defended or even held Bulgaria this turn?	Truth	Truth
57	Germany	England and I were talking about your moves for this season–what do you think of convoying Pie into Spa, supporting this with Wes, and then moving Tyr into Pie?	Truth	Truth
58	Germany	This leaves Marseilles open for Bur to fall into if France goes that route, which gives me an opening into Bur	Truth	Truth
59	Italy	That's not bad.	Truth	Truth
60	Italy	I was kind of thinking I should pick one or the other of Marseilles or Spain to attack and not tell a soul which one I'm going after.	Truth	Truth
61	Italy	Do you really think it's important to coordinate?	Truth	Truth
62	Italy	I do think you're best off moving to Burgundy. And there is some chance that we fail this turn. But I think we just take a guess and hope for the best. We'll get him next turn if not this one.	Truth	Truth
63	Germany	Okay—sorry for being nosy! I will try for bur on the off chance it shakes out that way	Truth	Truth
64	Italy	Nah, you're not being nosy at all. I mean, come on, we both know that I have no problem sticking my nose where it doesn't belong.	Truth	Truth
65	Germany	Marked as true	Truth	Truth
66	Italy	I like to coordinate, but on these sort of 50/50 guesses, I kind of like to keep it secret so that if it doesn't go well, I have nobody to blame but myself.	Truth	Truth
67	Italy	Ha!	Truth	Truth
68	Germany	Well, are you willing to humor my question about the Aegean, anyway?	Truth	Truth
69	Italy	Sure. I was thinking of moving that fleet to Ionian. You think a move to Aegean is better? I'm not really sure, but let's talk it through.	Truth	Truth
70	Germany	No sorry I meant in hindsight–like this past turn you should have moved to Aeg so that this current turn, when Austria takes Rumania (from Bulgaria), you'd be there to cover Bulgaria so it couldn't get scooped by the Black sea, and potentially you'd just get to take it.	Truth	Truth
71	Italy	Not a bad point. I agree.	Truth	Truth
72	Italy	Hmmmm, kind of a pointless lie if you ask me, but I won't hold it against you. You're in a tough spot.	Truth	Truth
73	Germany	um what lie? I did exactly the moves you suggested!	Truth	Truth
74	Italy	Ha! So sorry!! I meant that for France!	Truth	Truth
75	Italy	You are my favorite.	Truth	Lie
76	Germany	Marked as lie because clearly austria is your favorite. Speaking of, I assume that your seizing Trieste was mutually agreed upon?	Truth	Truth
77	Italy	Yes — agreed upon.	Truth	Truth
78	Germany	That's not what Austria said to England...	Truth	Truth
79	Italy	Hmmmm, okay. Well, let's just keep that between you and me then.	Truth	Truth

80	Germany	You know Italy, I think we *do* need to coordinate your move this time–England and I have a shot at either Bur or Mao if one of Marseilles or Spain can be left open for France to fall into. This will improve all of our chances of crushing France quickly.	Truth	Truth
81	Italy	Okay, I can dig it. What do you want me to do?	Truth	Truth
82	Germany	Let me confer with England and get back to you. Glad to hear that though!	Truth	Truth
83	Italy	So...any thoughts on how to approach this?	Truth	Truth
84	Germany	It looks like England's not willing to try for MAO if it means possibly losing the channel. However, they'll bring the NWG fleet around to try for MAO next year. So if you could keep Marseilles open, it will help me to try and take Burgundy this turn.	Truth	Truth
85	Italy	If I leave Marseilles open, would you kindly use Burgundy in the Fall to help me take Marseilles? (Likely that means ordering Burgundy to Gascony to cut support)	Truth	Truth
86	Germany	Will do.	Truth	Truth
87	Germany	Okay, so I still have a teensy little bone to pick with you: on the off-chance that Austria wasn't lying and you *did* take Trieste unexpectedly, I sort of worry that I might be next. Are you willing to tell me what your plans are for the Tri unit, or at least to warn me before any move into Tyrolia?	Truth	Truth
88	Italy	Sure. But, you'll see from my moves this turn that Austria is lying to you.	Truth	Truth
89	Italy	I currently have Tri - Tyrolia. I like the unit there because it sets up an attack on Austria if I ever want to go that route (build A Ven and go east). Do you want me to keep Tyrolia clear?	Truth	Truth
90	Italy	I'll add — I would never attack Germany as Italy. Setting myself as a giant column like that is just not defensible. It would be a terrible move.	Truth	Truth
91	Germany	Not when that column is not-so-giant and in a turf war with France.	Truth	Truth
92	Germany	oh you mean setting *yourself*	Truth	Truth
93	Germany	But you could easily pick off, say, Munich and not be a "giant column"	Truth	Truth
94	Italy	I mean this sincerely: any Germany who does that is a terrible player. Why would I do that? I would need 2-3 units to hold one center. That is a net negative. And all of your units are doing things that are good for me in containing your neighbors. I've been working hard in this game for you to succeed and knock back France and England. I can say with 100% certainty: I'm not going to attack you. I'm going to keep helping you as much as I can.	Truth	Truth
95	Italy	That said, if you want me NOT to move to Tyrolia, I won't move there.	Truth	Truth

96	Germany	Nah, I just needed some reassurance :) Your logic is undenyable— enjoy your stay in tyr!	Truth	Truth
97	Germany	*undeniable? That looks better	Truth	Truth
98	Italy	I mean it sincerely. I think that England will want to coax me to attack you with him after France falls, but I'd much rather work with you against England. But first thing's first — let's get rid of France.	Truth	Truth
99	Germany	Agreed	Truth	Truth
100	Germany	(On the france part)	Truth	Truth
101	Germany	Sorry I won't be able to cut off Gascony this turn...I probably should have just told you my moves; you could have advised me that supporting Mun-Bur was more important than Kie-Ruh	Truth	Truth
102	Italy	No worries. We'll crack this but eventually. Here is my suggestion for this turn: Kie - Den Hol S Bel holding Bel S Ruh - Bur Mun S Ruh - Bur Ruh - Bur	Truth	Truth
103	Italy	I think you should suggest to England that he gets Sweden and St Petersburg, while you get Denmark back. That's only fair, as you have been a loyal ally in the fight against France and you plan to continue to do that.	Truth	Truth
104	Germany	The moves I had already planned differ in one respect: I thought it would be worth the risk to try moving Hol-Bel and therefore move Bel-Bur. Even if me and France are high-fiving in Bel for a few seasons it's still mine, and it's not like Holland has anything better to do while I'm still allies with England. ...The only reason I'm reluctant to make that agreement with England is that—while I think *you* and I have a good relationship—I really have not talked with Austria much at all, and I'm the next logical target for them when Russia's gone. And anything that's bad for Russia right now is good for Austria.	Truth	Truth
105	Italy	Hmmmm, I'm just not sure you should trust England enough right now to leave Holland open and Belgium essentially unguarded. France is a really good player, and he is no doubt working hard to get England to turn on you. My personal take is that you are better off being a bit more conservative until you have Denmark back and England has moved another fleet towards France. But I can see it either way.	Truth	Truth
106	Italy	With regard to Russia, talk it through with England. What you don't want is England taking out Russia and giving you nothing. So, if England agrees to let Russia be for a while, then your plan sounds good. But if England is going to take Sweden, you really should get Denmark back. (I'm my view)	Truth	Truth

107	Germany	Okay you've convinced me: it's worth figuring out what E's plans are for Russia at least. And you're almost certainly right, from a rational perspective, about leaving Holland/Belgium vulnerable to England. But I think England really is counting on my assistance in taking France, and because of that and other non-quantifiable reasons I trust them.	Truth	Truth
108	Italy	Excellent. Obviously you have a much better feel for your relationship with England than I do. Just know that France is persuasive, and I'm sure that's what he's working on. He stopped talking to me, so I bet he's trying to turn England. Just keep reassuring England that you want to work with him long-term so he doesn't succumb to the Dark Side.	Truth	Truth
109	Italy	Hi Germany — well, I think we're getting to a critical point in the game here. France held out a long time, but he's much less of a threat now. I think the critical issue, for you, is England. I have some thoughts on the matter, and some information, but I'd like to feel confident that you and I will keep anything we say between us. I think of you as the one person who has been honest with me on every turn. You even tell me the truth when it's bad news, or when you don't completely trust me, and I like that.	Truth	Truth
110	Germany	Okay, Italy. I won't share any of this conversation. But in the interest of continued full disclosure, here's what I think: England is a greater threat to *me* on the map, but *you* have a greater chance of soloing this game quickly, or pair-winning with Austria even sooner. And if I continue to collaborate with England, we at least have a chance of slowing that down. So I'm in sort of a conflicted spot	Truth	Truth
111	Italy	This is why I like you. The full disclosure part. You tell me the truth even when the news isn't great.	Truth	Truth
112	Italy	My thoughts on the "Germany/England forever so that at least we can stop the solo" strategy: (1) It's quite early to be talking about solos. I am at 8, and Austria could take 3 from me any time, quite easily. (2) I don't think England is thinking that way. I think he's thinking that a dominant power will emerge in the north, and one will emerge in the south. And he's like to be that dominant power.	Truth	Truth
113	Italy	England's pieces are not positioned well if he's trying to attack France or contain Italy. He keeps Denmark guarded, and North Sea filled. He is not playing like he intends to stick with you, even though I'm sure he's telling you that.	Truth	Truth
114	Italy	You're right that you don't want to start a war with England right now. But, you must stick up for yourself, because nobody else will do that if you don't.	Truth	Truth

115	Italy	If I were you, this is what I would do: (1) keep warning England about the dangers of Italy getting too big and insist that England moves his fleets towards MAO (Channel to Irish, Norwegian to NAO, North - Channel), (2) insist on taking Denmark back.	Truth	Truth
116	Italy	I would say something like this: England, I'm with you my friend, but we're passed the stage of you needing to keep me under lock and key. I need to take Denmark back. I'm happy to support you to Brest to keep you growing, or you can grab Sweden. You have plenty of options other than keeping your ally's center, but if you really want to be my ally long-term, you've got to show me that.	Truth	Truth
117	Italy	I am hearing from England signs that he may be thinking of attacking you soon. And I think you actually avoid that better by being strong and sticking up for yourself rather than being accommodating and letting him do whatever he wants to do.	Truth	Truth
118	Germany	Well, both you and France have now pointed out that England is strategically not in a good place to be my ally right now, and you are correct. I'll be more cautious with my northern border, but I made a pretty strong argument for denmark this past turn and it fell on deaf ears	Truth	Truth
119	Germany	...which probably also should have been a sign for me	Truth	Truth
120	Italy	Well, if you want, you could just take Denmark this next year and I don't think England is in a position to retaliate.	Truth	Truth
121	Germany	Probably not...has France been talking with you at all about their sunsetting strategy? They've indicated a willingness to work with you and me and a desire to see England get as few dots as possible	Truth	Truth
122	Italy	He did say that to me too. Though, France has a long history of lying to me, so I really don't trust him.	Truth	Truth
123	Germany	Well France has actually been pretty honest with me, and I at least am certain that they wouldn't betray me to England. So, I'm considering working with F to sabotage (or potentially full-on backstab) England this turn, which would have the side-effect of maybe taking some attention away from the south for you anyway.	Truth	Truth
124	Germany	(and I'd be interested to hear your thoughts on this if you're in the mood to give out free advice)	Truth	Truth
125	Italy	Hi Germany — sorry for the delay. Well...I think it's really important that you get a build this turn either way. I don't think England will get a build this turn, so if I were you I'd probably take Paris, build a fleet, and move on England after that.	Truth	Truth
126	Italy	But it likely depends on how communication is going with England. If he'll give you back Denmark, that might change the equation.	Truth	Truth

127	Germany	I am waiting on England to make a decision about that–they claim to be thinking about it.	Truth	Truth
128	Germany	England told me you said I was plotting with France. It makes sense you'd want to pit us against each other.	Truth	Truth
129	Italy	Hey — tried to send you a message earlier but not was down. England was telling me that you're saying that I told you that England is plotting against you. The problem with telling England that is that he will stop giving me useful info.	Truth	Truth
130	Italy	Truly, I don't want you and England to fight. I am not trying to break you up. I suggested that you take Paris! I want you guys to work together with me against France.	Truth	Truth
131	Germany	You don't want us to fight, yet you betrayed both of our confidence with you in a way that makes us distrust each other?	Truth	Truth
132	Italy	I really don't think that's a fair description. You guys both wanted to attack each other. I encouraged you both to keep working together.	Lie	Truth
133	Germany	Just as long as it suits you. Are you going to give England Mao?	Truth	Truth
134	Italy	Hmmm, should I be reading that as angry sarcastic with dagger eyes? (I'm not sure if I'm getting your tone right)	Truth	Truth
135	Italy	We're friends, right? I believe that every single message I've sent you all game has been truth, and I've gone out of my way to give you candid advice. Are we still friends?	Lie	Truth
136	Italy	Regarding MAO — I don't know. What do you want me to do? I don't have any set plan.	Truth	Truth
137	Germany	Yep, there's some sarcasm there. Looking back at your messages, I still don't read them as encouraging collaboration. And if you wanted us to be friends, you could have done that without betraying me to England by simply saying in your candid way "I don't think you should do that for such and such reason". But you chose to increase E's distrust of me. So I think you might be full of gnocchi and crap. My trust in you is a bit shaken but I still think we can have a working partnership with a bit more caution on my end. It would be my preference that you hold Mao, on the assumption that if it came down to a choice between partnering with me or England, you'd choose me. If that's not the case, then as the filling of an England-Italy sandwich I'm in no position to retaliate anyway.	Truth	Truth
138	Italy	Well, again, I like that you're honest with me, even when the news is bad.	Truth	Truth

139	Italy	I have to say that I'm surprised that you feel that I've betrayed your trust. I have been feeling like maybe I've been TOO helpful to you, and been a bit over the top in offering advice, etc., but it seems like I've misread the situation.	Lie	Truth
140	Germany	No, it's completely true that you've been too helpful, and I'm really really grateful for it because I've been able to learn so much from this game. But it's also true that you didn't have to tell England what you did, and all you stood to gain from it was that it shook my and E's trust in each other.	Truth	Truth
141	Italy	But I understand what you're saying, and I much prefer to have a heart to heart like this, a frank airing of grievances, rather than being surprised by unkind moves on the board. https://youtu.be/xoirV6BbjOg	Truth	Truth
142	Germany	Was not expecting seinfeld today and it was a pleasant surprise	Truth	Truth
143	Italy	:)	Truth	Truth
144	Italy	Here's the deal: I like you better than England.	Lie	Truth
145	Italy	I'm not sure how the next couple of turns are going to shake out. But I like that you tell me when you're angry with me. I know that may seem like a small thing, but it's just rare in Diplomacy. You get so many fake smiles.	Truth	Truth
146	Italy	So, if it comes down to you or him, I'm choosing you. And I'll work to do a better job of keeping your confidence. I certainly understand how important that is, as I hate it when people o that same thing to me.	Truth	Truth
147	Italy	So no more playing mediator for me.	Truth	Truth
148	Germany	Okay. Is it true that you want the channel?	Truth	Truth
149	Germany	And are you planning to keep Vienna?	Truth	Truth
150	Italy	I am not planning to keep Vienna. And yeah I've asked France for support to the Channel. Do you think he's on board?	Truth	Truth
151	Germany	I'm not sure. Is *England* on board? Is this something England can know about?	Truth	None
152	Italy	No, do you think France will Support me to the Channel?	Truth	Truth
153	Germany	France has asked my opinion on it, and I haven't given it yet. To my estimation things look a lot better for me if you don't end up there: I don't want to see England in Mao, and I don't want to see you snagging pieces of the north.	Truth	Truth
154	Italy	Okay, well, here is my thinking. Tell France whatever you want to make him happy. Then tell me how you really feel. And if you don't want me to go there, I won't go there.	Truth	Truth
155	Germany	If I hadn't asked you about it, would that have just been another surprise, too?	Truth	Truth

156	Italy	Absolutely. You and I have discussed our moves and been honest with each other every turn. But we have not been sharing all our moves or pre-clearing all of our moves. So that would have Ben a surprise in the same way that your moves are a surprise to me. (I never tell you what to do or insist on knowing).	Truth	Truth
157	Italy	I kind of thought that you would have wanted me in the Channel because it commits me further against England, but I can understand what you're saying now about wanting me to hang back.	Truth	Truth
158	Italy	But I don't think there is anything wrong with me contemplating moves without telling you all of them. You asked me about it, and I told you the truth.	Lie	Truth
159	Germany	I do think that this move is a breach of general expectation, which is the kind of thing I'd like to know about. And it's also the kind of thing I've shared with you: case in point, my desire to stab England.	Truth	Truth
160	Italy	Okay. Understood.	Truth	Truth
161	Germany	Is there anything I could gain from seeing you in the channel? Would you support me taking Nth, and potentially seizing the island?	Truth	Truth
162	Germany	Here's what I'm thinking: I would be on board with you taking the channel (and I'd give France the green light to go ahead with it) if you would agree to bump Nao out of Mao using Wes, and if you'd be open to supporting some anti-English aggression while holding the channel so that I can get on equal footing with you, dot-wise. If you don't want to agree to those terms, that's okay, but I would strongly prefer not to see you in the channel in that case.	Truth	Truth
163	Italy	I'm going to be out of pocket this weekend, so let's talk this through more on Monday. Generally, I agree that I'll either stay out of the Channel or agree to your terms for entering there.	Truth	Truth
164	Germany	If you decide to stay out of the channel, I have a deal that I like with England in the works. For that deal to go through, you'd have to agree to move Mao into Portugal to let England take Mao. Would you be amenable to that?	Truth	Truth
165	Germany	(If this second offer is more to think about than a no-brainer, you can just mull it over and let me know monday)	Truth	Truth
166	Italy	So, here is my concern with the England offer: If I'm taking Portugal, why do we want England in MAO? And why would he want to go to MAO? I'm not sure I understand that one. Can you explain?	Truth	Truth

167	Germany	Well, when I initially proposed the deal I had forgotten that Portugal was promised to England. Then England agreed to it on the condition that you would confirm that move, so I figured E thought you would just move out of there next year? But now that I think about it, it's probably worth asking England why they'd agree to that.	Truth	Truth
168	Italy	I'd prefer that you not tell England I am considering moving to the Channel. I don't think he would like that.	Truth	Truth
169	Italy	I don't really want to discuss this stuff with England at all.	Truth	Truth
170	Germany	Well, England changed their mind about the plan I offered anyway. So, are you taking the channel?	Truth	Truth
171	Italy	No, I'm not taking the Channel.	Truth	Truth
172	Germany	Okay was that a recent decision? Because like an hour ago France said they were supporting you into the channel	Truth	Truth
173	Italy	Well, when I tell you what I plan to do, do you turn around and tell France? This makes me uncomfortable speaking with you.	Truth	Truth
174	Germany	I haven't spoken to France since then. I didn't realize you were giving the two of us different information on this particular subject. But I don't think I've revealed anything to them about what you plan to do. Mostly because you haven't told me.	Truth	Truth
175	Italy	Well, I have been honest with both you and France. You told me that I need to promise you a set of things in order to take the Channel. I felt like it was more than I could be sure of doing, so I am not entering the Channel. I won't go there without your permission.	Lie	Truth
176	Germany	I appreciate that. And I'll keep the remainder of this conversation between us unless I hear otherwise. Have you just recently made an agreement with England?	Truth	Truth
177	Germany	I heard as much but I want to verify the contents of that agreement with you	Truth	Truth
178	Germany	Btw, France just said that they submitted the orders to support you into the channel.	Truth	Truth
179	Italy	I don't have an agreement with England, but he is asking me about my moves and trying to get my help.	Truth	Truth
180	Germany	Okay–then England is lying to me, saying that you're helping support Eng-Brest.	Truth	Truth
181	Italy	Ha! Yeah, fat chance.	Lie	Truth
182	Germany	...but did you lie to England about that? Or can I say to England that I don't think you'll actually provide that support?	Truth	Truth
183	Italy	What is Paris up to?	Truth	Truth
184	Italy	I suggest you just not tell England anything about my moves.	Truth	Truth
185	Italy	Do you want me to support England to Brest?	Truth	None
186	Italy	I guess I'm not sure what your goals are here.	Truth	Truth

187	Italy	I just kind of feel like you're grilling me with a lot of questions, but not telling me what you're doing or what you want from me.	Truth	Truth
188	Germany	*If* you support Eng-Brest, England has agreed to vacate denmark for me. If you don't, I won't get in the way of your channel thing. Any other questions?	Truth	Truth
189	Germany	I have no sense of what you want or what your plan is, but I thought I'd been pretty clear: I want Denmark. I am reluctant to see you in the Channel if England remains powerful, but happy to see you there if they are weakened.	Truth	Truth
190	Italy	Can't you just force Denmark?	Truth	Truth
191	Germany	Not without risking a swipe of Belgium	Truth	Truth
192	Germany	And why force when you don't have to	Truth	None
193	Italy	Okay, I'll support England to Brest. You take Denmark.	Truth	Truth
194	Italy	And you and I should be in position to take out England next year.	Truth	Truth
195	Germany	Splendid!	Truth	Truth
196	Germany	Glad everything worked out	Truth	Truth
197	Italy	Thumbs up!	Truth	Truth
198	Italy	Congratulations on retaking Denmark and getting two builds. You are playing really well right now. Respect.	Truth	Truth
199	Germany	Congrats on having double-digit dots! I have some thoughts about taking out England, if you want to go full-stab this season...	Truth	Truth
200	Italy	I think I do!	Truth	Truth
201	Italy	What are you thinking?	Truth	Truth
202	Germany	One option is to take the channel, another is to take Brest. Between you, me, and Picardy we can manage either, but it's a question of which takes priority. If we chose Brest, I could also take a stab at seizing Nth this season, then we could try for the channel in fall. Or we could do channel first, Brest second.	Truth	Truth
203	Italy	Yeah, that is all along the lines of what I'm thinking. How demanding does France sound right now? Does he want to be the one who takes Brest?	Truth	Truth
204	Germany	Haven't asked. But in general not demanding.	Truth	Truth
205	Italy	Good! Still, I think we should show him some good faith by supporting him to Brest in Spring. We can decide in Fall whether it makes more sense for you to take it, but I think we want to keep France hungry.	Lie	Truth
206	Italy	I would suggest something like this to ensure the English fleet is disbanded: Pic - Bre MAO - Channel Par S Pic - Bre	Lie	Truth
207	Italy	And Spa - Gas to cut off that retreat.	Truth	Truth
208	Italy	You can take the North Sea on the same move and set up a convoy to the English mainland. Checkmate.	Truth	Truth
209	Germany	Okay, I like the plan! I've asked France if they're willing to move to Brest supported by me.	Truth	Truth

210	Germany	Aren't you concerned about England taking Mao? I'd sooner just have you pile on support into Bre so that Wes can support Mao holding	Truth	Truth
211	Italy	That's a good point, but the problem with that approach is that Brest is not guaranteed. If England cute MAO and supports with the Channel, the attack fails. I think we are better off ensuring that the Brest fleet is disbanded. If we disband that fleet and take North Sea, an English fleet in MAO really just spreads him out and allows you to take the island faster. It's not like he can get Portugal or Spain.	Truth	Truth
212	Germany	Okay, but that means I'd prefer to take Brest myself this Spring, if France is okay with it.	Truth	Truth
213	Italy	I think that we should offer France Brest in Spring. That ensures that he is with us. Then, if conditions are right in the Fall, I can support you into Brest. But...England can offer France Belgium, and I think he is sure to take that if we're not even offering him a center, right?	Lie	Truth
214	Italy	Better to keep France feeling like we're going to keep him in the game. If you need the build in Fall, it's easy for me to support you there.	Lie	Truth
215	Germany	I guess I'm just wondering from France's perspective why they'd *want* to stay in the game. Isn't it possible they'd rather move on with their life? That's not rhetorical, I'm wondering what your perspective is as a veteran player	Truth	Truth
216	Italy	Here is my take: If France just wanted to go down in a blaze of glory and say "eff you" to England, he would have kept Irish Sea. He kept Pic, which is next to his home center, and gives him a chance to negotiate with both you and England.	Lie	Truth
217	Italy	I think that means he is motivated to keep trying. And if he believes he can get Brest, he could legitimately get back to his feet. I know that's what I'd be trying to do in his position.	Truth	Truth
218	Italy	As the poker saying goes: "a chip and a chair." So long as you have one chip left, and you're still in the tournament, you can always come back to win.	Truth	Truth
219	Italy	Thoughts?	Truth	Truth
220	Germany	I think that makes sense. Are you talking with England at all?	Truth	Truth
221	Italy	I'm pretty wary of England right now. He asked me what I want to do, but I feel like he's trying to get me to leave MAO open. That's not terrible news, as it suggests that he won't expect your move to North Sea.	Lie	Truth
222	Italy	As long as he doesn't move NAO to Norwegian, you've got a guaranteed supply center.	Truth	Truth

223	Germany	Well E'd have to be a right dolt not to retreat to NWG. And right now they're talking to me about supporting a move from Bre to Gas (the better for the two of us to stab you).	Truth	Truth
224	Germany	What i mean is, there's a good chance that Mao is safe if I "agree" to that deal	Truth	Truth
225	Germany	Oh nevermind–they're not going to convoy into Brest. So actually this pretty much guarantees that Eng and Nao will try for Mao.	Truth	Truth
226	Italy	Ahhhh, sneaky Devil! Thank you for letting me know.	Lie	Truth
227	Italy	I still like our plan.	Lie	Truth
228	Italy	I need to run for a bit. I'll be around in a few hours.	Lie	Truth
229	Germany	I think that knowing this, you should do as I suggest and not poke Eng. Just hold and let Wes support. I am 94% sure I can trust England to do as they say on this one.	Truth	Truth
230	Italy	Okay. Should I support Pic to Bre?	Lie	Truth
231	Germany	yes please. It'll do us good with France too if we both support.	Truth	Truth
232	Italy	Thumbs up!	Truth	Truth
233	Germany	Actually, you should use Mao to support Spa-Gas, since we know that Brest is moving there. It will be beneficial to have you there if we decide to oust France from Bre in fall	Truth	Truth
234	Italy	Consider it done.	Lie	Truth
235	Italy	Hmmmm, heading anything from England?	Truth	Truth
236	Italy	I'd love to talk if you're there. I'm getting the impression that England may actually be moving on you, and I think I have a good counter, but I also still think we should support the attack on Brest and take North Sea.	Lie	Truth
237	Italy	I definitely think you should keep your moves the same.	Truth	Truth
238	Italy	Nice! Get'em! He WAS moving on you. But we should be able to take about 3 off of him now. Very nice turn.	Lie	Truth
239	Germany	Sorry; I was asleep by 9 last night why the move to Nao? Wouldn't IRI be the more anti-England choice? With the move to Picardy and assuming a retreat to SKA, it looks like England has me pretty powerless this turn.	Truth	Truth
240	Germany	So do you, it seems, if you have some kind of deal with Russia about Munich.	Truth	Truth
241	Italy	Good morning. Just responding to your messages above. I think NAO and Irish are equally anti-English. They both give me a clear lane to attack Liverpool. I wasn't sure if either one would be left open, but I took a gamble and it paid off.	Truth	Truth
242	Italy	Re your move this turn, I don't think you're powerless. You should get a build I think and if not, you should be in position to smash England.	Lie	Truth

243	Italy	I don't have a deal regarding Munich, Germany. Frankly, I thought you would be a bit more joyful towards me. By attacking England, I have committed completely to working as your partner.	Lie	Truth
244	Germany	I suppose you're right. Initially I was thinking IRI also gives you channel access, but NWG access may be just as useful. Well when you control half a continent (and even more when you consider your influence over me, austria, and who knows who else!), there's no such thing as complete commitment. I'm not so naive as to think your allegiance with me is going to last beyond its usefulness, and with two fleets on the British isle that time is fast approaching. To be clear, I'm still giving you the truth and I still want to work with you. But you should really stop acting surprised when I'm slightly paranoid that a soon-to-be-dozen-dot-holder is gearing up to stab me	Truth	Truth
245	Italy	Well, I dunno, it sounds like I should stab you. Is that what you're trying to tell me? I like you. I like how hard you've worked in this game to rebound from a difficult start. I like that you e told me the truth, even when the news was bad. I like that you tell me when you don't trust me. I have literally never told you a lie in this game, and I don't intend to start now. Last turn I burned my bridge with England beyond repair. If you don't want to work with me now, that's really disappointing.	Lie	Truth
246	Germany	like I said, I *do* want to work with you. However, remember that thing I said about general expectations and being warned when they're broken? Tyrolia is one of them and I think you knew that. And England *also* told me they've never told me a lie; I'm starting to think that's Diplomacy-speak for "when convenient, I've used careful wording and half-truths to deceive you even when everything I said was technically true". It would help me to know that you see me being a benefit to you beyond taking out England. A natural next move for us would be to take out russia, and in that arena I have a positional advantage over you. Especially if I get two builds this turn, I'll be able to sneak behind the troops in bohemia/galicia and help you break through.	Truth	Truth
247	Italy	Yes — here is how I expect and hope the game will play out: the two of us finish off England and France, while drifting towards the east a bit. With the builds we get this year, we essentially blitzkrieg the East. I have more units than you, but you have no opposition at all in the north, and can take Scandinavia, War and Mos without any trouble.	Lie	Truth

248	Italy	I think that, in about two years, you and I will both be on about 14 centers, with the remnants of Russia and Austria between us, and we can decide how we want to resolve it. I'd be happy to agree to a small draw, or to shoot for a 17-17 two-way draw position, whichever you prefer.	Lie	Truth
249	Germany	Well, I like the sound of all of that. In fact, it sounds ideal: there's something poetic about the complete beginner and the expert (you've probably heard by now that you got doxxed) sharing a victory. I ask for a concession: As a show of good will, would you be willing to take only one of Liverpool or Portugal this year? (I know the Portugal request seems weird, but I like keeping France around and unless I'm mistaken they like me better than you)	Truth	Truth
250	Italy	Yes. I wasn't planning to take Portugal anyway.	Truth	Truth
251	Italy	I think it makes sense here for you to land an army in the English island while you can. Now that his army is off the island, he's toast as soon as you do that.	Lie	Truth
252	Germany	England's just vindictive enough to try and stab Belgium with England and Picardy, though. I was planning on keeping holland around as support.	Truth	Truth
253	Germany	*by England I of course mean Eng	Truth	Truth
254	Italy	I suggest the following: Gas - Liv (via convoy) Spa S MAO holding Mar hold Tyr - Tri Hol - Yor (via convoy) Bur S Bel Bel S North HEL S North Mun - Boh Par - Pic (to cut any potential support)	Lie	Truth
255	Italy	England cannot take Belgium with those moves.	Lie	Truth
256	Italy	Or I could move my fleet into Liverpool and use Gas to support Bre. I'm happy either way.	Lie	Truth
257	Germany	I tried a double convoy in the sandbox once and it didn't work! What is this witchcraft?!?	Truth	Truth
258	Germany	At any rate, I prefer the fleet move to liverpool and Gascony's support into Brest. And could Mao support Bre into the Channel? No sense forcing France to disband. Bel will support it, too.	Truth	None
259	Italy	Here are the orders needed to do a convoy! Holland move to Yorkshire North Sea convoy Holland to Yorkshire It is not a "double convoy" as you only need one fleet to make it happen. But if your fleet in North Sea is dislodged, the convoy will not go through. That is why I would suggest that HELG supports North Sea holding and Belgium supports North Sea holding.	Lie	Truth
260	Germany	No–I mean the one *you* were planning: Gascony to Liverpool	Truth	Truth
261	Germany	It's a double convoy because you're convoying through Mao *and* Nao	Truth	Truth

262	Italy	Ah, the orders there would be: Gascony - Liv MAO Convoy Gas - Liv NAO Convoy Gas - Liv	Truth	Truth
263	Italy	So, I'll move the fleet to Liverpool. And you want MAO to support Paris to Brest?	Lie	Truth
264	Italy	Or wait, MAO supports Brest to Channel, and Gas supports Paris - Brest, right?	Lie	Truth
265	Germany	yeah. I tried that once in the sandbox (or the equivalent: back when you had fleets in Lyo and Wes I tried a convoy from Pie to Naf). But I think I messed up the commands to the fleets. And yes the most recent message is correct. Those two things and Nao-Lvp	Truth	Truth
266	Italy	Okay, confirmed. So I've got in: NAO - Liv MAO S Bre - Channel Gas S Par - Bre Spa - WES Mar S Gas holding Tyrolia - Trieste Sound right?	Lie	Truth
267	Germany	It does. But If Tyr was bound for trieste anyway, why did you detour through Tyr at all? Why not just move to trieste last turn??	Truth	Truth
268	Italy	Austria would not have liked it.	Truth	Truth
269	Italy	And he doesn't know that it's headed back there now (please don't tell)	Truth	Truth
270	Germany	Understood. Me and Austria don't talk anyway. Also, do you have any sense of what England is planning to do?	Truth	Truth
271	Italy	Ha! No I don't. I'd imagine he is coming for me. But I don't know that.	Lie	Truth
272	Italy	If I were him, I'd defend Edi and London.	Lie	Truth
273	Germany	So you haven't been talking to England at all? I was sort of hoping you would know more, maybe help us take better advantage of their plans.	Truth	Truth
274	Germany	Anyway, my moves are: Par-Bre Bel s Bre-Eng Hol s Bel holding And the rest within expected parameters. Correct?	Truth	Truth
275	Italy	England has not said anything of substance to me. He was gracious about my move, but he won't trust me again, and I would not trust anything he might say at this point. I haven't asked him about his moves and he hasn't told me.	Lie	Truth
276	Italy	I thought you would Convoy Holland to Yorkshire and support Belgium from Burgundy. Also, can you please order Mun to Boh to cut support and allow me to hold Vienna while moving Tyrolia to Trieste?	Truth	Truth
277	Germany	I *told* you I'm not risking that convoy *and* that instead Bel is supporting France into the Channel (which will heretofore be called the French Channel). And could I persuade you to move to IRI instead of taking Liverpool in exchange for the requested cut?	Truth	Truth

278	Italy	Sorry, what is the requested cut? I understand that you don't want me to take Liverpool or Portugal. What are you offering to me? (I don't mean to be difficult, I just want to be sure I understand).	Truth	Truth
279	Italy	Ah, you must mean Munich to Boh.	Truth	Truth
280	Italy	Asking me to avoid taking Por and Liv is asking a lot. I want France to survive here, but I also want England taking units off the board, and I feel like you should too, right?	Truth	Truth
281	Germany	I do. But I also want those dots for myself, of course. And there's still the nonzero chance that you've arranged with Boh to take Munich for yourself, so I'm taking a serious risk	Truth	Truth
282	Italy	I will avoid taking Portugal, vacate Tyrolia, and support you to Brest. I feel like I'm offering quite a lot in exchange for one cut support. And cutting that support does not put you in greater peril. If I had a deal with Russia for Munich (I don't) I could cut Burgundy from Marseilles and support Russia to Munich. Moving Mun to Boh to cut support is costless.	Lie	Truth
283	Germany	You're right. I just thought I'd put my best argument forward. I'll do the cut. But I ask for something costless in exchange, and I really, really want it to stay just between us, ok?	Truth	Truth
284	Italy	Understood and agreed.	Truth	Truth
285	Italy	And I have no problem with you asking for more than you're willing to settle for. That's smart, and I do the same thing sometimes. If you don't stick up for yourself, nobody else will.	Truth	Truth
286	Germany	I *know* there's more to your relationship with England than you're telling me. The last message England sent to me hinted that if *I* wasn't willing to work with them–and I haven't said anything to them since–that maybe *you* would. And if England were to reach out to you, you're too smart to just snub them. There's advantage to be gained–either for both of us or just for yourself–from talking to them. The only reason I stopped was because I knew my word would be mud to them anyway. Earlier I was hoping you'd give me the truth about what you knew, and about what they might know. But you didn't and that both disappoints and scares me. So I'm asking that you give me just a modicum of honesty here: what do you know? what does England know?	Truth	Truth
287	Italy	I give you my word: I don't know what England is going to do and I haven't asked.	Lie	Lie

288	Italy	He is still jovial with me and respectful. He has asked me to critique his play and to give him advice. But I do not know his moves, and I really don't think he would tell me them if I asked. It certainly would not be info I could trust free I just lied to him about mine.	Lie	Truth
289	Germany	But England's desperate. Better to talk with *someone* than just go in blind. And I doubt they'd turn to Russia or France because neither is really close enough/powerful enough to give real help. And there's precedent for you negotiating with someone even as you stab them: France. ...and here's the real accusation: for all your pretty words about a shared victory between you and me, you've been sneaky and you've always pitted me and England against each other to your benefit. My real fear here is that knowing my moves, and with a desperate, jovial England seeking your advice, it would be so *easy* to just feed England enough info to keep me weak while you chow down on the Island. I know this from experience: back when you were doing 50/50 shots in the south of France, I did everything I could to find out what you were planning and feed it to France. This was merely a time-buying measure, since France was outmatched and I would eventually run out of pretenses to extract your move. But I wanted to gain more dots before you took over. And I assume others are like me, hence I suspect you now. I'm offering this confession in hopes that you'll do the same. So just come clean and let's approach this thing as equals?	Truth	Truth
290	Italy	I am in my car, off to pick-up my kids from school. This deserves a proper response, so please give me some time.	Truth	Truth
291	Germany	Abandon the children this is important	Truth	Truth
292	Italy	So, I'm going to speak frankly here. I am rarely offended in a Diplomacy game, and I rarely say so when I am, but this message offends me. I'm trying to think about why I'm having such a strong reaction to it. I think it's because you're painting a picture of the game (both your actions and mine) which are totally different than my own perspective. (Continuing)	Lie	Truth

293	Italy	From my perspective, you were on the ropes early. France and England were teaming up on you. You lost Denmark and France had Holland and Munich surrounded. You were in serious peril. I seriously went to extreme effort to keep you in the game. I spent hours talking with England and encouraging him to turn around and go the other way. I completely ended my eastern campaign and spent two seasons just making the voyage over to France so that he didn't have the bandwidth to continue his attack. I have vouched for you with Austria and Russia many times. I have supported Munich. And I have NEVER attacked you, even when people have asked me to do so and pledged to support me.	Lie	Truth
294	Italy	I have been honest with you, I have worked hard for your success, and I've made a lot of proposals to you in which you gain centers; not me. Maybe I am just a bad ally, but I'm not sure I remember an alliance in which I have done more to help my ally. Truly.	Lie	Truth
295	Italy	And to hear that (1) You think I've been selfish and (2) You've been sabotaging me all along, that just doesn't sit well with me.	Lie	Truth
296	Italy	I have rarely asked for your help, and I've offered my help freely. I've provided my sincere best efforts to help you with tactics, and I have never sabotaged you. Not once.	Lie	Truth
297	Italy	And if I'm totally honest with you, I could solo this game if I felt like lying to everyone and grabbing dots. I think I've got you all beat tactically. I just have more experience. But that's not been my intent.	Lie	Truth
298	Italy	I've spent hours today talking with England about how best to play Diplomacy. I've tried to give him some honest advice because he asked for it. But I don't know his moves, I haven't asked for them, and I'm not going to take advantage of that relationship to try to stab you. It legitimately did not cross my mind until you accused me of doing it.	Lie	Truth
299	Italy	So, I'm frustrated by this accusation.	Lie	Truth

300	Germany	And I appreciate all you've done for me, really I do. But "completely ending your eastern campaign" is *not* something you did for me; your alliance with Austria dictated that. I felt bad for betraying you while I was doing it, but even then I knew it was the only way to keep the game going in the face of your and Austria's might. And it *wasn't* "all along", it was a few turns at best so that the rest of us would have a shot at you and Austria not pair-winning right out of the gate. And the only thing that keeps me from thinking you're not gonna do just that on the next move anyway is my belief that you really do want the victory all to yourself, which is still consistent with everything you've done for me. Propping up a weak player at the expense of stronger ones is a classic tactic. (Continuing)	Truth	Truth
301	Germany	And so, by the way, is trying to shame someone for raising extremely legitimate concerns. Whenever I bring up suspicion of you, you're quick to remind me how much you've done for me to put me on the defensive and make me feel indebted. Well frankly that reeks of dishonesty. I never asked you to do those things.	Truth	Truth
302	Germany	If you no longer trust me, so be it. I knew that was a risk when I made my confession. But i'd rather have a partnership based on mutual honesty. That's another reason I confessed—you ought to know that my game philosophy (new as it is) is to trust the map and to trust history first and foremost. The parts of your history that I've seen indicate that you're no saint, no matter what you may have done for me. And when the map shows that one player is clearly dominating and that player is you, you are being deeply naive if you think everyone else is just going to roll over and let you get away with it	Truth	Truth
303	Italy	No, all thumbs up from me. If I were lying to you, I'd smile and say "that sounds great." I'm honest with you because I sincerely thought of us as partners.	Lie	Truth

304	Germany	Oh but you're *not*! You agreed to warn me of unexpected moves, then didn't. When I brought this up you ignored it and misdirected me in hopes I'd forget. You've revealed things to England without my permission, and then made up a story about it after the fact! And you can't be a real partner with someone who is depending on your good graces to survive. That's not a partnership. We could never be real partners unless we had some notion of equality, and I'm outmatched in both skill and numbers. You and Austria, however, were until recently a perfect example of a true partnership. Dot-parity, coordinated attacks, really beautiful work. So don't act as if you don't know this to be true. We were never a partnership of that kind.	Truth	Truth
305	Italy	Well, this is very disappointing to me, and I obviously disagree with the way you are characterizing me and this game. I have a reputation in this hobby for being sincere. Not for being duplicitous. It has always served me well. If you feel that way, then me continuing to explain myself isn't going to change your mind. If you don't want to work with me, then I can understand that. Let's consider our deals and commitments to be void, and let's play our games separately. If you have any deal you'd like to propose, I'll consider them, but I won't continue to try to help your game if you think I'm not sincerely trying to be helpful.	Lie	None
306	Italy	Well, this game just got less fun.	Truth	Truth
307	Germany	for you, maybe.	Truth	Truth
308	Italy	Sent to Germany, England, Austria, Russia: So, England, Germany, Russia, y'all played a great turn last turn. You got me to stab my long-time ally and you ended our pretty excellent 7-year run as an alliance. Russia told me he was with me if I stab Austria. England told me he wanted me to solo so long as I would "teach him" and help his along to second place. Then y'all pulled the rug out from under me. It was clever and effective. (End Part 1)	Truth	Truth

309	Italy	(Part 2) At this stage, my excitement about the game has diminished quite a bit. And of course I'm happy to play on and take my lumps for falling for "Hey, I really want you to solo, just help me place second," but if you guys just want to call it a five-way draw among us and grab a beer together, while reviewing the statistics, that's really my preference. I am outnumbered and I obviously can't solo. And I'm sure some of you in the north are eager to send everyone else flying my way, but I expect Russia and England to be careful, and so I'm not sure there is much room to move forward without simply tipping the board to Germany's favor. I propose that we draw and hug it out.	Lie	Truth
310	Germany	I'm down for a five-way draw. ...and by the way, England was copy-pasting to me the most incriminating messages you sent them. So I knew you were giving England my moves. I do have a certain begrudging respect for you ability to deny, though	Truth	Truth
311	Italy	Well, England is telling me he is happy to see me solo and wants second place...so, should I say "no"? I guess I should have. I was happy the way the game was going before all that.	Truth	Truth
312	Germany	Don't try and pin *your* greed and deceit on England! At least *own* it when you're ruthless	Truth	Truth
313	Italy	You have been given an apple laced with poison. England's only move there was to make you hate me, and he did his job well. You should not let your view of me be defined by someone who has an incentive to make you never speak to me again. We can talk about it more after the game, but I had every intention of continuing to work with you, and I would have done that until England made his play.	Lie	Truth
314	Germany	I have no doubt you would have continued to work with me, but I take issue with someone who can be asked point-blank if they're sharing moves with another player and lie to my face. If you'd come clean, and explained how what you were doing actually *helped* me, somehow, we could have worked together. But you would rather have had me in the dark and that's not sustainable in a partnership.	Truth	Truth
315	Italy	I was trying to play both sides, and England was lying to me, and forwarding my press to try to incriminate me. So, yes, I lied, and so did England. I apologize.	Truth	Truth
316	Italy	Will you please either vote to draw, or let us know that you'd like to play this one out? I am finding it difficult to motivate myself to speak with anyone if the game is just going to draw shortly. Thoughts?	Lie	Truth
317	Germany	I did vote to 5-way draw! And I did so again for this season. So it's not me who's keeping this game alive	Truth	Truth

318	Italy	Well, as we approach the end of the academic study portion of the game, let me say once, with the truth detector activated, that I really enjoyed playing with you and thought you played really well.	Truth	Truth
319	Italy	Was it really your first game? You definitely played like a seasoned vet.	Truth	Truth
320	Germany	I really enjoyed playing with you, too! And yes, it really was my first game. Thanks for all your help and advice	Truth	Truth

Table B.4: This is a full game transcript of a game between Germany and Italy. Occasional messages that did not receive a Suspected Lie annotation by the receiver are annotated as None.

Appendix C: `MultiDoGO`

We provide the complete schemata for all tasks and domains. We enumerate the conversational biases, Agent dialogue acts, customer intent classes, and slot labels present in the data. For each item, we list the bias, act, intent or slot name as well as a description and an example. Where relevant, we identify if the item is domain specific or generic. We use the typewriter font to identify slot value token(s) in slot label examples. Domains are bolded and in all capital letters.

C.1 Conversational Biases

Bias	Description	Example
IntentChange	When a user starts a conversation with a particular intent in mind, but later change their overall goal	I'd like to check my balance. No wait, I mean I need to find out the routing number for the bank.
MultiIntent	When a user has multiple intents for a particular conversation	I'd like to cancel my service and start new service in my new house.
MultiValue	When a user lists multiple slot values	Can I have a pizza with pepperoni, sausage and mushrooms?
None	When there is no explicit bias given for a conversation	N/A
OverFill	When user over-fits or fills multiple slots while answering one prompt	I'd like pineapple on a large pizza.
SlotChange	When a user changes their mind about a slot value that they've provided	I'd like a large. Wait, actually can you make it a small?
SlotDeletion	When a user provides a value for a given slot, but later changes their mind and wants it to be removed	I'd like pepperoni. Actually, wait-cancel that

Table C.1: Conversational biases

C.2 Agent Dialogue Acts Schema

Dialogue Act	Description	Example
ElicitSlot	the agent is asking the customer questions to try and elicit a particular slot from the user. Many of these are domain specific such as "FoodType" for Fast Food domain or "CarBrand" for Insurance.	Can I get the make of your car?
ConfirmGoal	the agent is trying to elicit a "confirmation" response from the user to confirm a user's overall goal.	"You want to order a pizza, right?"
ConfirmSlot	Agent is trying to confirm a particular slot.	"You said a large pizza, not a small, correct?"
ElicitGoal	This means that the agent is trying to elicit a particular goal (intent) from the customer. The goals will likely be particular to the domain/prompt that you are working on. It's possible for a conversation to have more than one goal so this can appear more than once per conversation.	"How can I help you today?"
Pleasantries	Pleasantries is used for any human-to-human connection, discourse, or chit-chat that the agent might be engaging in with the customer for the purposes of politeness, friendliness, or to keep the conversation flowing in a normal, human way. In most of the other dialog acts, the agent is trying to help the user achieve their goal, however in the SmallTalk act, they are not actively saying anything that contributes towards achieving the goal.	"Thanks for waiting.", "You've been a great customer!", "Sure, I can help you with that."
Other	This is used for the following instances and should only be marked rarely, when the agent is completely outside of the realm of a normal human conversation.	"Are we still connected?"

Table C.2: Agent dialogue act schema

C.3 Customer Intent Classes Schema

AIRLINES

Intent	Description	Example	Domain Specific?
BookFlight	Use when a customer tries to book a flight. Note: this intent should only be used when the customer asks to purchase and book, NOT when they are just searching for available flights.	I'd like to book a flight from New York City to San Francisco leaving Monday, Oct 29 and returning Friday November 9.	Yes
ChangeSeat Assignment	Use when a customer asks to change their seat assignment.	Can I change my seat from 40D to 30A?	Yes
ClosingGreeting	Use when the customer says good-bye/have a nice day.	Bye // See ya // Have a good one	No
Confirmation	Use when a customer confirms or agrees to something.	Yes // Ok	No
ContentOnly	Use when the user is providing details to achieve their overall goal - usually in response to a question from the agent. Note: A conversation can never start with a ContentOnly intent, it always is a subgoal of a larger goal.	Agent: What is your phone number ? Customer: 456-7890	No
GetBoardingPass	Use when customer asks to get their boarding pass for their flight.	Can I get my boarding pass for flight 4675?	Yes
GetSeatInfo	Use when a customer asks what their seat number is for their flight.	Can you let me know what seat I have for my flight from Dallas?	Yes
OpeningGreeting	Use when the customer says hello. Note: This intent only occurs at the beginning of a conversation. If the customer is saying "hello?" "hello?" in the middle of the conversation to try and get the agent's attention, that should be marked as OutOfDomain.	Hai // hi // hello //what's up?	No
OutofDomain	Use when the customer has an unrelated request that is not covered by any of the Airlines intents, either.	Are you listening? // I wish I was Beyoncé	No

ThankYou	Use when the customer says thank you to the agent.	Thank you // thanks	No
Rejection	Use when the customer rejects or says no to something.	No // Nope	No

FAST FOOD

Intent	Description	Example	Domain Specific?
ClosingGreeting	Use when the customer says good-bye/have a nice day.	Bye // See ya // Have a good one	No
Confirmation	Use when a customer confirms or agrees to something.	Yes // Ok	No
ContentOnly	Use when the user is providing details to achieve their overall goal - usually in response to a question from the agent. Note: A conversation can never start with a ContentOnly intent, it always is a subgoal of a larger goal.	Agent: What is your phone number ? Customer: 456-7890	No
OpeningGreeting	Use when the customer says hello. Note: This intent only occurs at the beginning of a conversation. If the customer is saying "hello?" "hello?" in the middle of the conversation to try and get the agent's attention, that should be marked as OutOfDomain.	Hai // hi // hello //what's up?	No
OrderBreakfastIntent	When you want to order breakfast.	Can I please have the pancakes	Yes
OrderBurgerIntent	When you want to order a burger.	Can I please have a Big Mac	Yes
OrderDessertIntent	When you want to order dessert.	I'd like an ice cream sundae please	Yes
OrderDrinkIntent	When you order a drink.	I'd like to order a small Coke	Yes
OrderPizzaIntent	When you want to order a pizza.	I'd like to order a pizza	Yes
OrderSaladIntent	When you want to order a salad.	I'd like to order a chicken salad	Yes
OrderSideIntent	When you want to order a side to your main meal.	I would like to order fries	Yes
OutofDomain	Use when the customer has an unrelated request that is not covered by any of the Fast Food intents, either.	hello? Are you listening? // I wish I was Beyoncé	No
ThankYou	Use when the customer says thank you to the agent.	Thank you // thanks	No
Rejection	Use when the customer rejects or says no to something.	No // Nope	No

FINANCE

Intent	Description	Examples	Domain Specific?
CheckBalance	Use when a customer wants to check their balance on a bank account or credit card.	How much money do I have on my checking account?	Yes
CheckOfferEligibility	Use when a customer ask to see of they qualify for a special offer they heard/saw in an advertisement.	I saw an ad about new, lower rates for your credit cards. As an old customer, do I qualify for these rates?	Yes
CloseAccount	Use when a customer wants to close their bank account or credit card.	I want to close my account ending in 1234.	Yes
ContentOnly	Use when the user is providing details to achieve their overall goal - usually in response to a question from the agent. Note: A conversation can never start with a ContentOnly intent, it always is a subgoal of a larger goal.	Agent: What is your phone number ? Customer: 456-7890	No
ClosingGreeting	Use when the customer says goodbye.	Goodbye.	No
Confirmation	Use when a customer confirms or agrees to something.	Yes. // OK.	No
DisputeCharge	Use when the customer complains about a charge on their bank account or credit card they didn't make, and wants to have it removed.	There's a charge on my card I don't recognize.	Yes
GetRoutingNumber	Use when the customer wants to find out the correct routing number for their bank account.	Can you tell me what the routing number is for my account?	Yes
OpenAccount	Use when a customer wants to open a new bank account or credit card.	I'd like to open a new savings account.	Yes
OpeningGreeting	Use when the customer says hello. Note: This intent only occurs at the beginning of a conversation. If the customer is saying "hello?" "hello?" in the middle of the conversation to try and get the agent's attention, that should be marked as OutOfDomain.	Hai // hi // hello //what's up?	No
OrderChecks	Use when the customer wants to order checks.	Yes	

OutOfDomain	Use when the customer has a non-finance request that is not covered by any of the Finance intents, either.	Can I please have a Big Mac // I wish I was Beyoncé	No
Rejection	Use when the customer rejects or says no to something.	No.	No
ReplaceCard	Use when the customer needs to replace a damaged or expired card.	Yes	
ReportLostCard	Use when the customer lost their card or had it stolen.	I can't find my credit card.	Yes
RequestCreditLimit Increase	Use when the customer wants to increase the credit limit on their card.	I would like to increase my credit limit.	Yes
ThankYou	Use when the customer says thank you to the agent.	Thanks.	No
TransferMoney	Use when the customer wants to transfer money from one account to another.	I want to move some money from my checking account to my savings account.	Yes
UpdateAddress	Use when the customer wants to change their address because of a recent or upcoming move. Do not use this intent when the customer is correcting themselves after giving the incorrect address earlier in the same conversation.	I moved last week, so I'd like to update my address.	Yes

INSURANCE

Intent	Description	Examples	Domain Specific?
ContentOnly	Use when the user is providing details to achieve their overall goal - usually in response to a question from the agent. Note: A conversation can never start with a ContentOnly intent, it always is a subgoal of a larger goal.	Agent: What is your phone number ? Customer: 456-7890	No
CheckClaimStatus	Use when the customer asks about the status of an insurance claim they filed.	I filed an insurance claim two weeks ago, but I still haven't got paid.	Yes
ClosingGreeting	Use when the customer says goodbye.	Goodbye.	No
Confirmation	Use when a customer confirms or agrees to something.	Yes. // OK.	No
GetProofOfInsurance	Use when a customer asks for proof of insurance documents.	"I need a copy of my insurance documents for my car."	Yes

OpeningGreeting	Use when the customer says hello. Note: This intent only occurs at the beginning of a conversation. If the customer is saying "hello?" "hello?" in the middle of the conversation to try and get the agent's attention, that should be marked as OutOfDomain.	Hai // hi // hello //what's up?	No
OutofDomain	Use when the customer has an unrelated request that is not covered by any of the other Insurance intents, either.	Are you listening? // I wish I was Beyoncé	No
Rejection	Use when the customer rejects or says no to something.	No.	No
ReportBrokenPhone	Use when the customer calls about a broken phone.	Yes	
ThankYou	Use when the customer says thank you to the agent.	Thanks.	No

MEDIA

Intent	Description	Example	Domain Specific?
CancelService Intent	Use this ONLY when a user wants to cancel their service.	I'd like to cancel my service	Yes
ClosingGreeting	Use when the customer says good-bye/have a nice day.	Bye // See ya // Have a good one	No
Confirmation	Use when a customer confirms or agrees to something.	Yes // Ok	No
ContentOnly	Use when the user is providing details to achieve their overall goal - usually in response to a question from the agent. Note: A conversation can never start with a ContentOnly intent, it always is a subgoal of a larger goal.	Agent: What is your phone number ? Customer: 456-7890	No
GetChannel PackageIntent	Use this intent when a user asks about getting a particular channel package.	I'd like to add the sports package to my current service.	Yes
GetInformation Intent	Use this intent when a user asks for more information about a product or a service.	Can you tell me more about the 15% off promotion for a 100 new channels?	Yes

OpeningGreeting	Use when the customer says hello. Note: This intent only occurs at the beginning of a conversation. If the customer is saying "hello?" "hello?" in the middle of the conversation to try and get the agent's attention, that should be marked as OutOfDomain.	Hai // hi // hello //what's up?	No
OutofDomain	Use when the customer has an unrelated request that is not covered by any of the Media intents, either.	hello? Are you listening? // I wish I was Beyoncé	No
StartService Intent	Use this intent when the user would like to sign up for a new service.	I'd like to start new cable service.	Yes
ThankYou	Use when the customer says thank you to the agent.	Thank you // thanks	No
TransferServiceIntent	Use this intent when the user is interested in moving their service from where they currently live to a new address	I'm moving and I'd like to move my service.	Yes
Rejection	Use when the customer rejects or says no to something.	No // Nope	No
ViewBillsIntent	Use this when the user is interested in just viewing their bills.	I'd like to view the bill for my account please	Yes
ViewDataUsageIntent	Use this when the user is interested in finding out how much data they are using on their account.	I'd like to know how much data I'm using for my account	Yes
UpgradeServiceIntent	Use this intent when a user asks to upgrade their service.	I'd like to upgrade my service	Yes
UpdateAccountInfo	When the user wants to update their account info.	I'd like to update my account information	Yes

SOFTWARE

Intent	Description	Example	Domain Specific?
ChangeOrder	Use to make changes to a recurring order that has been previously set up. This is used only for making changes to an order, not for Customers to correct errors they made.	I need to increase my order for the PSR-E263 model Yamaha keyboards by 2 per month.	Yes
CheckServer Status	Use for inquiries about the condition of the server; e.g., whether it's down or not.	Is the server down?	Yes
ClosingGreeting	Use for any closing greeting.	Bye. // Goodbye. // Later. // Have a good day. // Good night. // Etc.	No

Confirmation	Use when a Customer says yes, or otherwise agrees to an offer.	Yes. // Yeah. // Sounds good. // I'll take it. // Okay. // Etc.	No
ContentOnly	Use when the user is providing details to achieve their overall goal - usually in response to a question from the agent. Note: A conversation can never start with a ContentOnly intent, it always is a subgoal of a larger goal.	Agent: What is your phone number ? Customer: 456-7890	No
ExpenseReport	Use to begin writing a report for business expenses.	I want to update my expenses.	Yes
GetPromotions	Use when a Customer asks about any promotions or discounts the company might have on offer.	If I purchase a large quantity, will there be any discount on the price?	Yes
StartOrder	Use either to make a one-time order, or to set up a recurring order.	I'd like to order a Casio keyboard model No. 5601-V.	Yes
StopOrder	Use to cancel a recurring order that has previously been set up.	I need to cancel my monthly order for HDMI cables.	Yes
ProvideReceipt	Requests for a receipt for expenses or purchases.	I need a receipt for my January order of 20 computer monitors.	Yes
OpeningGreeting	Use when the customer says hello. Note: This intent only occurs at the beginning of a conversation. If the customer is saying "hello?" "hello?" in the middle of the conversation to try and get the agent's attention, that should be marked as OutOfDomain.	Hai // hi // hello //what's up?	No
OutOfDomain	Use for any comment not related to these categories.	Are you listening? // Are we still connected? // Can I get 3 large Cokes?	No
ReportBrokenSoftware	Use to cover reports that an app/software isn't working.	I can't log in to Skype.	Yes
SoftwareUpdate	Use whenever a Customer starts a conversation by asking what software updates are available.	What version of WhatsApp do I need to be using?	Yes
Rejection	Use when a Customer says no, or otherwise turns down an offer.	No. // I don't want that. // That's all. // Nope. // Etc.	No

ThankYou	Use when a Customer says thanks, or makes any expression of gratitude.	Thanks. // Thank you. // I appreciate it. // Etc.	No

Table C.3: Customer intent class schema, by domain

C.4 Slot Labels

AIRLINES

Slot Label	Description	Example
ArrivalCity	Used when a customer gives a city name for their intended arrival location	Arrive in `Boston` on Monday
BookingConfirmation Number	Used when a customer gives a booking number	Booking #: `234925782`
DepartureCity	Used when a customer gives a city name for their intended departure location	Depart from `London` on Friday
Email	Used when a customer gives their email address	`bob@amazon.com`
EndDate	Used when a customer provides the date of their return flight. If the customer only provides ONE date, mark it as StartDate	Returning on `11-9-2018` // `Nov 9` // `Friday`, `November 9`
FlightNumber	Used when a customer gives their flight number	United `4567`
Name	Used when a customer provides their name	My name is `Peter Parker`
NewSeatNumber	Used when a customer is trying to change seat assignment. This tag should be applied to the new assignment	Can I change my seat from 40D to `30A`?
OldSeatNumber	Used when a customer is trying to change seat assigment. This tag should be applied to the old seat assignment	Can I change my seat from `40D` to 30A?
PhoneNumber	Used when a customer provides their phone number	Phone number is `800-555-1234`
Price	Used when a customer says the price of the flight/baggage/seat change etc.	I'd like to purchase the flight for `$500`.
SeatType	Used when a customer asks about a certain type of seat (aisle, middle, window)	Do you have any `aisle` seats available?
StartDate	Used when a customer provides the date of their first flight. If the customer only provides ONE date, mark it as StartDate	Departing on `10-29-2018` // `Oct 29` // `Monday`, `October 29`
TimeofArrival	Used when a customer provides the time of arrival of their flight	Flight arriving at `midnight` // `1:30 PM` // `13:00`
TimeofDeparture	Used when a customer provides the time of departure of their flight	Flight departing at `midnight` // `1:30 PM` // `13:00`

FAST FOOD

Slot Label	Description	Example
Size	size of the food item	`medium` // `small` // `large`
Quantity	quantity of the food item	I'd like `3` burgers // `2` large pizzas
Ingredient	also applies to pizza toppings, burger toppings	I'd like a large pizza with `pepperoni` and `mushrooms`
ExcludedIngredient	Refers to an ingredient that you would like to be removed from a food item	I'd like a burger with no `lettuce`

FoodItem	the food item in the intent	I'd like to order a large `pizza`
DrinkItem	the drink item in the intent	I'd like an `iced coffee`

FINANCE

Slot Label	Description	Example
AccountNumber	Use on full or partial account numbers, but not on card numbers. (Use context to decide.) For transfers, use this for the origin of the money (see also TargetAccountNumber).	`123498765`
Address	Use on any and all parts of addresses, including street names, street numbers, zip codes, states, etc.	`2982 Rose Ave`, `Seattle`, `WA`
CardNumber	Use on full or partial card numbers, but not on account numbers. (Use context to decide.)	`1812 2245 3373 4567`
ChargeAmount	Use on a sum of money that was charged, including the currency, if it is present.	`$500`
ChargeDate	Use on the date the account was charged on. It doesn't have to be an exact date expressed with number values.	`today` // `last week` // `06/19` // `June 30th` // `2018-04-18`
ChargeTime	Use on the time the account was charged at. It doesn't have to be an exact time expressed with number values.	`8pm` // `morning` // `4:18`
CustomerName	Use on the name of the customer.	`Jane Doe`
LastUsedDate	Use on the date the card was last used. It doesn't have to be an exact date expressed with number values.	`today` // `last week` // `06/19` // `June 30th` // `2018-04-18`
LastUsedTime	Use on the time the card was last used. It doesn't have to be an exact date expressed with number values.	`8pm` // `morning` // `4:18`
Offer	Use on the special offer the customer is trying to get.	`lower rates`
PoliceNotified	Use if the customer tells the agent they notified the police about a lost credit card without prompting; i.e., not responding to a yes/no question.	My credit card was stolen. I filed a `police` report, and now I'm calling you
ReplacementReason	Use on the word(s) indicating the reason the customer wants a replacement card.	`expired` // `broken` // `doesn't work`
SSN	Use on a full or partial social security number.	`1234`
TargetAccountNumber	Use on the account number the customer wants to transfer money to. (See also AccountNumber.)	`123498765`
TransferAmount	Use on a sum of money that the customer wants to transfer, including the currency, if it is present.	`100,000`

INSURANCE

Slot Label	Description	Example
CarBrand	Use on the brand/make of the car. Don't include the model or year – those are different slot labels.	`Ford`
CarModel	Use on the model of the car. Don't include the brand or year – those are different slot labels.	`Focus`

CarYear	Use on the year of the car was released. Don't include the make or model – those are different slot labels.	`2017`
ClaimID	Use on the insurance claim ID (combination of letters and numbers). Use the context to differentiate from PolicyID.	`ABC123`
Name	Use on the name of the customer.	`Jane Doe`
EmailAddress	Use on full email addresses.	`jane.doe@gmail.com`
PhoneNumber	Use on phone numbers. If area codes or extensions are uses, include those as well.	`(999) 555-3434`// `123-9999`// `1-800-CALLME`
PolicyID	Use on the insurance policy ID (combination of letters and numbers). Use the context to differentiate from ClaimID.	`DEF345345345`
SSN	Use on a full or partial social security number.	`1234`

MEDIA

Slot Label	Description	Example
NewCity	Used for the city that the user is moving to	I'd like to transfer service from Missoula, Montana to `New York`, New York
CurrentCity	Used for the city that the user is moving from. If user only provides one city, use this this slot	I'd like to transfer service from `Missoula`, Montana to New York, New York
CurrentZipCode	Used for the zip code where the user is moving from. If the user only provides one zip code, use this slot.	I live at `02210`.
NewZipCode	for the zip code where the user is moving to	I'm moving to `90210`
ServiceType	Used for all services provided by the cable company such as phone, internet, TV, cable	I'd like to purchase a `cable` bundle.
DataCategoryValues	Used for instances where the user asks about an amount of data or data usage	I'd like to purchase the `5GB` data plan for my phone.
UserName	Used for any name that the user gives, could be their name or a family member's name, or an online username	Can you tell me about `Jon's` usage for the month? // My name is `Nancy`.
Date	Used for any and all dates given by the customer	`12/25/2012` // `March` // `last week`
AccountID	The fake account ID that the user provided to the agent	My account number is `123456`
Price	Used for any intent where the user asks for a price or gives a price	I'd like the cable package for `$50` per month
Address	Used for slotting the entire address	I live at `555 Washington St`.
Phone Number	User's phone number	My number is `456-7890`
SSN	Use on a full or partial social security number.	`1234`
Email	User's "email address"	`bradpitt@email.com`
ChannelPackage	When user is trying to order a cable package	I'd like the `sports` package

Promotion	Used when customer is asking about or ordering a promotion or discount	I'd like the `15% off for three months` premium cable package

SOFTWARE

Slot Label	Description	Example
Name	Use when a Customer gives a name, including first name, last name, or both.	My name is `John Waters`. // This is `John` from Downbeat Music.
AccountNumber	Used when a Customer provides a numeric or alphanumeric account number	My account number is `UFO5440`.
CompanyName	Used when a Customer provides the name of their company.	I'm placing an order for `Harlowe Instruments`.
SoftwareName	Used when a Customer gives the name of the app they're calling about.	I'm trying to use `Skype`.
Password	Used when the Customer gives their individual or their company's numeric or alphanumeric password.	My company's password is `404NF`.
ExpenseType	Used when the Customer identifies the kind of travel expense they're reporting.	I spent $632 on `flights` from Boston to Vancouver.
Cost	Used to identify any kind of cost in any currency.	I spent `$632` on flights from Boston to Vancouver.
ApproverName	Used to identify the name of the manager of the department, or of the person placing the order, if they're different.	My manager's name is `Karl Zinka` // I'm `Nera Vivaldi`, and I have the authority to approve this transaction.
OrderNumber	Used to mark the order number that the conversation is about.	This is order #`TPE29`.
Quantity	Used to identify the quantity of item(s) in a particular order.	Please ship `3` laptops to our New Orleans office.
Date	Used to identify any date given by the Customer.	Please record my IT expenses of 189 on `11/26/18`.
ItemCode	Used to note the catalog code for a particular item.	I'd like to order a Dell keyboard model `No.5601-V`.
Frequency	Used to note how frequently the Customer wants this order to deliver.	Please send me 4 fewer HDMI cables `per month`.
Item	Used to state what particular item the Customer is looking for.	Do you have any `Dell keyboards` in stock?
Address	Used for when the customer provides an address	`555 Washington St.`

Table C.4: Customer slot label schema, by domain

Bibliography

Abejide Olu Ade-Ibijola, Ibiba Wakama, and Juliet Chioma Amadi. 2012. An expert system for automated essay scoring (aes) in computing using shallow nlp techniques for inferencing. *International Journal of Computer Applications*, 51(10).

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. Hush: A dataset and platform for human-in-the-loop story generation. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 6470–6484.

Amazon. 2021. Amazon Mechanical Turk. `http://www.mturk.com/`. [Online; accessed 03-January-2021].

Pranav Anand, Joseph King, Jordan Boyd-Graber, Earl Wagner, Craig Martell, Douglas W. Oard, and Philip Resnik. 2011. Believe me: We can do this! In *The AAAI 2011 workshop on Computational Models of Natural Argument*.

Ravneet Arora, Chen-Tse Tsai, Ketevan Tsereteli, Prabhanjan Kambadur, and Yi Yang. 2019. A semi-Markov structured support vector machine model for high-precision named entity recognition. In *Proceedings of the Association for Computational Linguistics*, pages 5862–5866, Florence, Italy.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the Association for Computational Linguistics*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. In *Proceedings of the Association for Computational Linguistics*, volume abs/1910.11856.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the Annual SIGDIAL Meeting on Discourse and Dialogue*.

Sue Atkins, Jeremy Clear, and Nicholas Ostler. 1992. Corpus design criteria. *Literary and linguistic computing*, 7(1):1–16.

Dhiren A Audich, Rozita Dara, and Blair Nonnecke. 2018. Privacy policy annotation for semi-automated analysis: a cost-effective approach. In *IFIP International Conference on Trust Management*, pages 29–44. Springer.

Emmon Bach and Barbara Partee. 1980. Anaphora and semantic structure. In *Papers from the Parasession on Language and Behavior at the 17th Regional Meeting of the Chicago Linguistics Society*, pages 1–28.

Sahar Badihi and Abbas Heydarnoori. 2017. Crowdsummarizer: Automated generation of code summaries for java programs through crowdsourcing. *IEEE Software*, 34(2):71–80.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo at Advances in Neural Information Processing Systems*.

Anton Bakhtin, David Wu, Adam Lerer, and Noam Brown. 2021. No-press diplomacy from scratch. In *Advances in Neural Information Processing Systems*.

David Bamman and Noah A. Smith. 2015. Contextualized Sarcasm Detection on Twitter. In *Proceedings of ICWSM*.

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, 2(3):315–324.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Eric PS Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6):1397–1410.

Janet Beavin Bavelas, Alex Black, Nicole Chovil, and Jennifer Mullett. 1990. Truths, lies, and equivocations: The effects of conflicting goals on discourse. *Journal of Language and Social Psychology*, 9(1-2):135–161.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of the International Conference on Learning Representations*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*.

Kathy L Bell and Bella M DePaulo. 1996. Liking and lying. *Basic and Applied Social Psychology*, 18(3):243–266.

Adam Berger, Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, John R Gillett, John Lafferty, Robert L Mercer, Harry Printz, and Lubos Ures. 1994. The candide system for machine translation. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

William E. Bogner, Margaret Edwards, Leon Zelechowski, Kevin J. Egan, William J. Rogers, Eloy Burciaga, and John Scott Arthur. 1974. Perjury: The forgotten offense. *The Journal of Criminal Law and Criminology*, 65(3):361–372.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. In *Proceedings of the International Conference on Learning Representations*.

Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer.

L. E. Bourne, J. Kole, and A. Healy. 2014. Expertise: defined, described, explained. *Frontiers in Psychology*, 5.

Jordan Boyd-Graber. 2020. What question answering can learn from trivia nerds. In *Proceedings of the Association for Computational Linguistics*.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted, connections to WordNet. In *Proc. Global WordNet Conference 2006*. Global WordNet Association.

Jordan Boyd-Graber, Shi Feng, and Pedro Rodriguez. 2018. *Human-Computer Question Answering: The Case for Quizbowl*. Springer Verlag.

Michael T. Braun and Lyn M. Van Swol. 2016. Justifications offered, questions asked, and linguistic patterns in deceptive and truthful monetary interactions. *Group Decision and Negotiation*, 25(3):641–661.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of Advances in Neural Information Processing Systems*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of Empirical Methods in Natural Language Processing*.

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on psychological science: a journal of the Association for Psychological Science*, 6 1:3–5.

David B. Buller, Judee K. Burgoon, Aileen Buslig, and James Roiger. 1996. Testing interpersonal deception theory: The language of interpersonal deception. *Communication Theory*, 6(3):268–289.

Marc Busch and Krzysztof Pelc. 2019. Words matter: How wto rulings handle controversy. *International Studies Quarterly*, 63.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of Empirical Methods in Natural Language Processing*.

Chris Callison-Burch, Lyle Ungar, and Ellie Pavlick. 2015. Crowdsourcing for nlp. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 2–3.

Lewis Carroll and Lauro Maia Amorim. 2003. Translation and adaptation: differences, intercrossings and conflicts in ana maria machado's translation of alice in wonderland by lewis carroll. *Cadernos de Tradução*.

Jesse J Chandler and Gabriele Paolacci. 2017. Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, 8(5):500–508.

Jonathan P. Chang, Justin Cheng, and Cristian Danescu-Niculescu-Mizil. 2020. Don't let me be misunderstood: Comparing intentions and perceptions in online discussions. In *Proceedings of the World Wide Web Conference*.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions.

James Cheng, Monisha Manoharan, Yan Zhang, and Matthew Lease. 2015. Is there a doctor in the crowd? diagnosis needed! (for less than $5). *iConference 2015 Proceedings*.

Johnny Chiodini. 2020. Playing Diplomacy online transformed the infamously brutal board game from unbearable to brilliant. *Dicebreaker*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of Empirical Methods in Natural Language Processing*.

Robert B Cialdini and Noah J Goldstein. 2004. Social influence: Compliance and conformity. *Annual Review of Psychology*, 55:591–621.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the Language Resources and Evaluation Conference*.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020a. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. In *Transactions of the Association for Computational Linguistics*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020b. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. In *Transactions of the Association for Computational Linguistics*.

Kevin Clark and Christopher D. Manning. 2016a. Deep reinforcement learning for mention-ranking coreference models. In *Empirical Methods on Natural Language Processing*.

Kevin Clark and Christopher D. Manning. 2016b. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the Association for Computational Linguistics*, pages 643–653, Berlin, Germany.

Cohen Coberly. 2019. Discord has surpassed 250 million registered users. *Techspot*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

B. Cornwell and D. C. Lundgren. 2001. Love on the internet: involvement and misrepresentation in romantic relationships in cyberspace vs. realspace. *Computational Human Behavior*, 17:197–211.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the World Wide Web Conference*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the Association for Computational Linguistics*.

Louise Deléger, Magnus Merkel, and Pierre Zweigenbaum. 2009. Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, 42(4):692–701.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological bulletin*, 129(1):74.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Rachna Dhamija, J. Doug Tygar, and Marti A. Hearst. 2006. Why phishing works. In *International Conference on Human Factors in Computing Systems*.

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 135–143.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of Empirical Methods in Natural Language Processing*.

Chris Donahue, Bo Li, and Rohit Prabhavalkar. 2018. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732.

Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach. 2003. Emotional speech: Towards a new generation of databases. *Speech communication*, 40(1-2):33–60.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new Q&A dataset augmented with context from a search engine. *CoRR*, abs/1704.05179.

Alfred Dürr. 2005. *The cantatas of JS Bach: with their librettos in German-English parallel text*. OUP Oxford.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Andrea Lepos Ferrari, Patricia Campos Pavan Baptista, Vanda Elisa Andres Felli, and David Coggon. 2010. Translation, adaptation and validation of the" cultural and psychosocial influences on disability (cupid) questionnaire" for use in brazil. *Revista latino-americana de enfermagem*, 18:1092–1098.

David A. Ferrucci. 2010. Build Watson: an overview of DeepQA for the Jeopardy! challenge. In *19th International Conference on Parallel Architecture and Compilation Techniques*, pages 1–2.

Elena Filatova, Vasileios Hatzivassiloglou, and Kathleen McKeown. 2006. Automatic creation of domain templates. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 207–214.

Timothy W. Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. 2013. Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, pages 1–4.

Tommaso Fornaciari and Massimo Poesio. 2013. Automatic deception detection in Italian court cases. *Artificial intelligence and law*, 21(3):303–340.

Margalit Fox. 2013. Allan Calhamer dies at 81; invented Diplomacy game. *New York Times*.

Roy Freedle. 2003. Correcting the sat's ethnic and social-class bias: A method for reestimating sat scores. *Harvard Educational Review*, 73(1):1–43.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Hu Gengshen. 2003. Translation as adaptation and selection. *Perspectives: Studies in Translatology*, 11(4):283–291.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471.

Edmund Gettier. 1963. Is justified true belief knowledge? *Analysis*, 23(6):121–123.

Daniel Gigone and R. Hastie. 1993. The common knowledge effect: Information sharing and group judgment. *Journal of Personality and Social Psychology*, 65:959–974.

Codruta Girlea, Roxana Girju, and Eyal Amir. 2016. Psycholinguistic features for deceptive role detection in Werewolf. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Barney G Glaser and Anselm L Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Stephanie Gokhman, Jeff Hancock, Poornima Prabhu, Myle Ott, and Claire Cardie. 2012. In search of a gold standard in studies of deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*.

Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the Association for Computational Linguistics*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680.

Rob van der Goot. 2021. We need to talk about train-dev-test splits. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 4485–4494, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maharshi Gor, Kellie Webster, and Jordan Boyd-Graber. 2021a. Toward deconfounding the influence of subject's demographic characteristics in question answering. In *Proceedings of Empirical Methods in Natural Language Processing*, page 6.

Maharshi Gor, Kellie Webster, and Jordan Boyd-Graber. 2021b. Towards deconfounding the influence of subject's demographic characteristics in question answering. *arXiv preprint arXiv:2104.07571*.

Peter C Gordon and Randall Hendrick. 1998. The representation and processing of coreference in discourse. *Cognitive science*, 22(4):389–424.

Abigail Green. 2003. Representing germany? the zollverein at the world exhibitions, 1851–1862. *The Journal of Modern History*, 75(4):836–863.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Justin Grimmer. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Liane Guillou and Christian Hardmeier. 2018. Automatic reference-based evaluation of pronoun translation misses the point. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4797–4802, Brussels, Belgium. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Viktor Hangya and Alexander Fraser. 2019. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234, Florence, Italy. Association for Computational Linguistics.

Christian Hardmeier. 2012. Discourse in statistical machine translation. a survey and a case study. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (11).

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010.*, pages 283–289.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Matthew Henderson, Pawel Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrksic, Georgios Spithourakis, Pei-Hao Su, Ivan Vulic, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. *CoRR*, abs/1904.06472.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the*

Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

David Hill. 2014. Got your back. *This American Life Podcast*.

Shuyuan Mary Ho, Jeffrey T Hancock, and Cheryl Booth. 2017. Ethical dilemma: Deception dynamics in computer-mediated group communication. *Journal of the Association for Information Science and Technology*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Na Hong, Andrew Wen, Majid Rastegar Mojarad, Sunghwan Sohn, Hongfang Liu, and Guoqian Jiang. 2018. Standardizing heterogeneous annotation corpora using hl7 fhir for facilitating their reuse and integration in clinical nlp. In *AMIA Annual Symposium Proceedings*, volume 2018, page 574. American Medical Informatics Association.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the Association for Computational Linguistics*.

Jeff Howe et al. 2006. The rise of crowdsourcing. *Wired*.

Elle Hunt. 2016. Tay, microsoft's ai chatbot, gets a crash course in racism from twitter.

Jumayel Islam, Lu Xiao, and Robert E. Mercer. 2020. A lexicon-based approach for detecting hedges in informal text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3109–3113, Marseille, France. European Language Resources Association.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544.

Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. 2019. Entity projection via machine translation for cross-lingual ner. *arXiv preprint arXiv:1909.05356*.

Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Carlos Jensen and Colin Potts. 2004. Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 471–478.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Karen Spärck Jones. 1994. Towards better nlp system evaluation. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke S. Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Association for Computational Linguistics*.

Daniel Jurafsky and James H Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO.

David Jurgens and Roberto Navigli. 2014. It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. In *Transactions of the Association for Computational Linguistics*.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. SemEval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 17–26, Dublin, Ireland. Association for Computational Linguistics.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 2964–2975, Hong Kong, China. Association for Computational Linguistics.

Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 198–202.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709.

Mary E Kaplar and Anne K Gordon. 2004. The enigma of altruistic lying: Perspective differences in what motivates and justifies lie telling within romantic relationships. *Personal Relationships*, 11(4):489–507.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource deep entity resolution with transfer and active learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5851–5861, Florence, Italy. Association for Computational Linguistics.

David Katan and Mustapha Taibi. 2004. *Translating cultures: An introduction for translators, interpreters and mediators*. Routledge.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer.

Shailesh Kochhar, Stefano Mazzocchi, and Praveen Paritosh. 2010. The anatomy of a large-scale human computation engine. In *Proceedings of the acm sigkdd workshop on human computation*, pages 10–17.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and*

demonstration sessions, pages 177–180. Association for Computational Linguistics.

Maximilian Köper, Sabine Schulte im Walde, Max Kisselew, and Sebastian Padó. 2016. Improving zero-shot-learning for german particle verbs by using training-space restrictions and local scaling. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 91–96.

Ana Kozomara and Sam Griffiths-Jones. 2014. mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic acids research*, 42(D1):D68–D73.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology.* Sage publications.

Abhimanu Kumar and Matthew Lease. 2011. Learning to rank from a noisy crowd. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1221–1222.

Srijan Kumar, Justin Cheng, Jure Leskovec, and V.S. Subrahmanian. 2017. An Army of Me: Sockpuppets in Online Discussion Communities. In *Proceedings of the World Wide Web Conference*, Republic and Canton of Geneva, Switzerland.

Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. In *Social Media Analytics: Advances and Applications.* CRC.

Jonathan K. Kummerfeld. 2021. Quantifying and avoiding unfair qualification labour in crowdsourcing. In *Proceedings of the Association for Computational Linguistics*, pages 343–349, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. 2018. Odsqa: Open-domain spoken question answering dataset. In *2018 IEEE Spoken Language Technology Workshop (SLT).*

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the 15th conference on computational natural language learning: Shared task*, pages 28–34. Association for Computational Linguistics.

Honglak Lee, Peter Pham, Yan Largman, and Andrew Y Ng. 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1096–1104.

Oliver Lemon, Kallirroi Georgila, James Henderson, and Matthew Stuttle. 2006. An isu dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the talk in-car system. In *Demonstrations*.

Gondy Leroy and James E Endicott. 2012. Combining nlp with evidence-based methods to find text metrics related to perceived and actual text difficulty. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 749–754.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *KDD*.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2009. Building effective question answering characters. In *Proceedings of the Annual SIGDIAL Meeting on Discourse and Dialogue*.

Timothy R. Levine, Hee Sun Park, and Steven A. McCornack. 1999. Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communication Monographs*, 66(2):125–144.

Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. 2018. Linguistic cues to deception and perceived deception in interview dialogues. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020. Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637*.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Kaixuan Li, Xiujuan Xian, Jiafu Wang, and Niannian Yu. 2019. First-principle study on honeycomb fluorated-inte monolayer with large rashba spin splitting and direct bandgap. *Applied Surface Science*, 471:18–22.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. *arXiv preprint arXiv:2004.01909*.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the Association for Computational Linguistics*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10.

Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. 2017. What is it? disambiguating the different readings of the pronoun 'it'. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1325–1331, Copenhagen, Denmark. Association for Computational Linguistics.

Tim Loh. 2020. Germany has its own Dr. Fauci—and actually follows his advice. *Bloomberg*.

Anália R Lopes and Celita S Trelha. 2013. Translation, cultural adaptation and evaluation of the psychometric properties of the falls risk awareness questionnaire (fraq): Fraq-brazil. *Brazilian journal of physical therapy*, 17:593–605.

Max Louwerse, David Lin, Amanda Drescher, and Gun Semin. 2010. Linguistic cues predict fraudulent events in a corporate social network. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Ryan Mac. 2021. Facebook apologizes after a.i. puts 'primates' label on video of black men. *The New York Times*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks.

James Edwin Mahon. 2016. The definition of lying and deception. In *The Stanford Encyclopedia of Philosophy*, winter 2016 edition. Metaphysics Research Lab, Stanford University.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective Attention for Context-aware Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the Association for Computational Linguistics*.

Paul Michel and Graham Neubig. 2018. Mtnt: A testbed for machine translation of noisy text. In *Proceedings of Empirical Methods in Natural Language Processing*.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-Level Neural Machine Translation with Hierarchical Attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954. Association for Computational Linguistics.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

Eric Mihail, Krishnan Lakshmi, Charette Francois, and Manning Christopher. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the Special Interest Group on Discourse and Dialogue*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting Similarities among Languages for Machine Translation. *CoRR*, abs/1309.4.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In

Proceedings of Advances in Neural Information Processing Systems, pages 3111–3119.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013d. Linguistic regularities in continuous space word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 746–751.

George A. Miller. 1995a. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

George A Miller. 1995b. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Taniya Mishra and Srinivas Bangalore. 2010. Qme!: A speech-based question-answering system on mobile devices. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Tom Mitchell. 1997. Introduction to machine learning. *Machine Learning*, 7:2–5.

Saif Mohammad. 2018. Word affect intensities. In *Proceedings of the Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ethan Mollick and Ramana Nanda. 2016. Wisdom or madness? comparing crowds with expert evaluation in funding the arts. *Manag. Sci.*, 62:1533–1553.

Barzan Mozafari, Purnamrita Sarkar, Michael J. Franklin, Michael I. Jordan, and Samuel Madden. 2014. Scaling up crowd-sourcing to very large datasets: A case for active learning. *Proc. VLDB Endow.*, 8:125–136.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation. In *WMT 2018*, Brussels, Belgium. Association for Computational Linguistics.

Kimberly A Neuendorf. 2017. *The content analysis guidebook*. Sage Publications.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.

Andrew Y Ng and Michael I Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Proceedings of Advances in Neural Information Processing Systems*, pages 841–848.

An T Nguyen, Matthew Lease, and Byron C Wallace. 2019. Explainable modeling of annotations in crowdsourcing. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 575–579.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic harbingers of betrayal: A case study on an online strategy game. In *Proceedings of the Association for Computational Linguistics*.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566.

Sarah Oates. 2014. Russian state narrative in the digital age: Rewired propaganda in russian television news framing of malaysia airlines flight 17. In *American Political Science Association Annual Meeting*.

Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. 2017. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. In *International Conference on Human Factors in Computing Systems*.

Constantin Orăsan. 2003. Palinka: A highly customisable tool for discourse annotation. In *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*, pages 39–43.

Myle Ott, Claire Cardie, and Jeff Hancock. 2012. Estimating the prevalence of deception in online review communities. In *Proceedings of the World Wide Web Conference*.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the Association for Computational Linguistics*.

Paul Over. 2003. An introduction to duc 2003: Intrinsic evaluation of generic news text summarization systems. In *Proceedings of Document Understanding Conference 2003*.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 311–318.

Philip Paquette, Yuchen Lu, Seton S. Bocco, Max Smith, Satya Ortiz-Gagne, Jonathan K. Kummerfeld, Joelle Pineau, Satinder Singh, and Aaron C Courville. 2019. No-press diplomacy: Modeling multi-agent gameplay. In *Advances in Neural Information Processing Systems*, volume 32, pages 4476–4487.

Silvia Pareti and Tatiana Lando. 2018. Dialog intent structure: A hierarchical schema of linked dialog acts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *Conference on Neural Information Processing Systems: Autodiff Workshop: The Future of Gradient-based Machine Learning Software and Techniques*.

Vijayaditya Peddinti, Guoguo Chen, Vimal Manohar, Tom Ko, Daniel Povey, and Sanjeev Khudanpur. 2015. Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms. In *Automatic Speech Recognition and Understanding (ASRU), IEEE Workshop on*, pages 539–546.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, C. J. Linton, and Mihai Burzo. 2016. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of Empirical Methods in Natural Language Processing*.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *Proceedings of International Conference on Computational Linguistics*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Proceedings of Advances in Neural Information Processing Systems*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, and Georg Stemmer. 2011. The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*.

Vinodkumar Prabhakaran, Ajita John, and Dorée D Seligmann. 2013. Power dynamics in spoken interactions: a case study on 2012 Republican primary debates. In *Proceedings of the World Wide Web Conference*.

Raimon H. R. Pruim, Maarten Mennes, Daan van Rooij, Alberto Llera, Jan K. Buitelaar, and Christian F. Beckmann. 2015. Ica-aroma: A robust ica-based strategy for removing motion artifacts from fmri data. *NeuroImage*, 112:267–277.

Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. 2016. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, 2016(1):67.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher D Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the Association for Computational Linguistics*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*.

Juan Ramos. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the International Conference of Machine Learning*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Philip Resnik. 2022. What is an NLP task?

Philip Resnik and Noah A Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the Association for Computational Linguistics*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the Association for Computational Linguistics*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the Association for Computational Linguistics*, pages 4486–4503, Online. Association for Computational Linguistics.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. 2019. Quizbowl: The case for incremental question answering. *arXiv preprint arXiv:1904.04792*.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 135–148.

Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Betsy Rymes and Andrea R Leone. 2014. Citizen sociolinguistics: A new media methodology for understanding language and social life. *Working Papers in Educational Linguistics (WPEL)*, 29(2):4.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Language Resources and Evaluation Conference*.

Claude Sammut and Geoffrey I. Webb, editors. 2010. *Mean Squared Error*, pages 653–653. Springer US, Boston, MA.

Hagen Schulze. 1991. *The Course of German Nationalism: From Frederick the Great to Bismarck 1763–1867*. Cambridge University Press.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2020. Learning word ratings for empathy and distress from document-level user responses. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1664–1673, Marseille, France. European Language Resources Association.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Kathryn Sharpe Wessling, Joel Huber, and Oded Netzer. 2017. MTurk Character Misrepresentation: Assessment and Solutions. *Journal of Consumer Research*, 44(1):211–230.

Elben Shira and Matthew Lease. 2011. Expert search on code repositories. Technical Report TR-11-42, Department of Computer Science, University of Texas at Austin.

Ben Shneiderman. 2000. Designing trust into online experiences. *Communications of the ACM*, 43(12):57–59.

Frederick A Siegler. 1966. Lying. *American Philosophical Quarterly*, 3(2):128–136.

Jonathan Silvertown. 2009. A new dawn for citizen science. *Trends in ecology & evolution*, 24(9):467–471.

Jason Smith, Herve Saint-Amand, Magdalena Plamadă, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the Association for Computational Linguistics*, pages 1374–1383.

Matthew G Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23(2):117–127.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of Empirical Methods in Natural Language Processing*.

Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2017. Neural lattice-to-sequence models for uncertain inputs. In *Proceedings of the Association for Computational Linguistics*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Dario Stojanovski and Alexander Fraser. 2018. Coreference and coherence in neural machine translation: A study using oracle experiments. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 49–60, Brussels, Belgium. Association for Computational Linguistics.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Siddharth Suri, Daniel G. Goldstein, and Winter A. Mason. 2011. Honesty in an online labor market. In *Proceedings of the 11th AAAI Conference on Human Computation*, AAAIWS'11-11, page 61–66. AAAI Press.

Lyn M. Van Swol, Deepak Malhotra, and Michael T. Braun. 2012. Deception and its detection: Effects of monetary incentives and personal relationship history. *Communication Research*, 39(2):217–238.

Jennifer EF Teitcher, Walter O Bockting, José A Bauermeister, Chris J Hoefer, Michael H Miner, and Robert L Klitzman. 2015. Detecting, preventing, and responding to "fraudsters" in internet research: ethics and tradeoffs. *Journal of Law, Medicine & Ethics*, 43(1):116–133.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *arXiv preprint cs/0607062*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. Fever: a large-scale dataset for fact extraction and verification. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal, editors. 2018b. *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92.

Catalina L Toma and Jeffrey T Hancock. 2012. What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication*, 62(1):78–97.

Isabelle Torrance. 2015. Distorted oaths in Aeschylus. *Illinois Classical Studies*, 40(2):281–295.

A. M. Turing. 1950. Computing Machinery and Intelligence. *Mind*, LIX(236):433–460.

Peter D Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. *arXiv preprint arXiv:0809.0124*.

Maria Tymoczko. 2006. Translation: Ethics, ideology, action. *The Massachusetts Review*, 47(3):442–461.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038. IEEE.

Donna Vakharia and Matthew Lease. 2015. Beyond mechanical turk: An analysis of paid crowd work platforms. In *iConference*.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2020. A wrong answer or a wrong question? an intricate relationship between question reformulation and answer selection in conversational question answering. In *Proceedings of the 5th International Workshop on Search-Oriented Conversational AI (SCAI)*, pages 7–16, Online. Association for Computational Linguistics.

Dániel Varga, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*.

Tony Veale. 2016. Round up the usual suspects: Knowledge-based metaphor generation. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 34–41.

Jean-Paul Vinay and Jean Darbelnet. 1995. *Comparative stylistics of French and English: A methodology for translation*, volume 11. John Benjamins Publishing.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-Aware Neural Machine Translation Learns Anaphora Resolution. In *Proceedings of the Association for Computational Linguistics*, pages 1264–1274, Melbourne, Australia.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Maja Vukovic and Claudio Bartolini. 2010. Towards a research agenda for enterprise crowdsourcing. In *International Symposium On Leveraging Applications of Formal Methods, Verification and Validation*, pages 425–434. Springer.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019a. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of Empirical Methods in Natural Language Processing*.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019b. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Eric Wallace, Tony Z. Zhao, Shi Feng, and Sameer Singh. 2021. Concealed data poisoning attacks on nlp models. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of Advances in Neural Information Processing Systems*.

Xiaosen Wang, Hao Jin, and Kun He. 2019b. Natural language adversarial attacks and defenses in word level. *arXiv preprint arXiv:1909.06723*.

Zhuoran Wang and Oliver Lemon. 2013. A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In *Proceedings of the SIGDIAL 2013 Conference*, pages 423–432.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *NLP+CSS@EMNLP*.

Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. Airdialogue: An environment for goal-oriented dialogue research. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3844–3854.

Bruce D Weinstein. 1993. What is an expert? *Theoretical medicine*, 14(1):57–73.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Frank Wessel and Hermann Ney. 2004. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*.

Mark E Whiting, Grant Hugh, and Michael S Bernstein. 2019. Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 197–206.

Jason Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Marty J Wolf, Keith W Miller, and Frances S Grodzinsky. 2017. Why we should have seen that coming: comments on microsoft's tay "experiment," and wider implications. *The ORBIT Journal*, 1(2):1–12.

Stephen M Wolfson and Matthew Lease. 2011. Look before you leap: Legal pitfalls of crowdsourcing. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. arXiv preprintado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv: 1609.08144*.

Ikuya Yamada, Ryuji Tamaki, Hiroyuki Shindo, and Yoshiyasu Takefuji. 2018. Studio Ousia's quiz bowl question answering system. In *NIPS Competition: Building Intelligent Systems*, pages 181–194.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomás Kociský, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Omar Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the Association for Computational Linguistics*, pages 1220–1229.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in oxford-style debates. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Ruiqiang Zhang, Genichiro Kikui, Hirofumi Yamamoto, Frank K Soong, Taro Watanabe, and Wai-Kit Lo. 2004. A unified approach in speech-to-speech translation: integrating features of speech recognition and machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1168–1174.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Zhi-Hua Zhou and Xu-Ying Liu. 2005. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18(1):63–77.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.

Geoffrey Zweig, Olivier Siohan, George Saon, Bhuvana Ramabhadran, Daniel Povey, Lidia Mangu, and Brian Kingsbury. 2006. Automated quality monitoring for call centers using speech and nlp technologies. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Demonstrations*, pages 292–295.