ABSTRACT

Title of dissertation: DETECTING FINE-GRAINED SEMANTIC DIVERGENCES TO IMPROVE TRANSLATION UNDERSTANDING ACROSS LANGUAGES

Eleftheria Briakou, Doctor of Philosophy, 2023

Dissertation directed by: Professor Marine Carpuat
Department of Computer Science

One of the core goals of Natural Language Processing (NLP) is to develop computational representations and methods to compare and contrast text meaning across languages. Such methods are essential to many NLP tasks, such as question answering and information retrieval. One of the limitations of those methods is the lack of sensitivity to detecting *fine-grained semantic divergences*, i.e., fine-meaning differences in sentences that overlap in content. Yet, such differences abound even in *parallel texts*, i.e., texts in two different languages that are typically perceived as exact translations of each other. Detecting such fine-grained semantic divergences across languages matters for machine translation systems, as they yield challenging training samples and for humans, who can benefit from a nuanced understanding of the source.

In this thesis, we focus on **detecting** fine-grained semantic divergences in parallel texts to improve machine and human translation understanding. In our first piece of work, we start by providing empirical evidence that such small meaning differences exist and can be reliably annotated both at a sentence and at a sub-sentential level. Then, we show that they can be automatically detected by fine-tuning large pre-trained language models without

supervision by learning to rank synthetic divergences of varying granularity. In our second piece of work, we turn to **analyzing the impact** of fine-grained divergences on Neural Machine Translation (NMT) training and show that they negatively impact several aspects of NMT outputs, e.g., translation quality and confidence. Based on these findings, we present two orthogonal approaches to **mitigating** the negative impact of divergences and improve machine translation quality: first, we introduce a divergent-aware NMT framework that models divergences at training time; second, we present generation-based approaches for revising divergences in mined parallel texts to make the corresponding references more equivalent in meaning.

After exploring how subtle meaning differences in parallel texts impact machine translation systems, we switch gears to understand how divergence detection can be used by humans directly. In our last piece of work, we extend our divergence detection methods to **explain** divergences from a human-centered perspective. We introduce a lightweight iterative algorithm that extracts contrastive phrasal highlights, i.e., highlights of segments indicating where divergences reside within bilingual texts, by explicitly formalizing the alignment between them. We show that our approach produces contrastive phrasal highlights that match human-provided rationales of divergences better than prior explainability approaches. Finally, based on extensive application-grounded evaluations, we show that contrastive phrasal highlights help bilingual speakers detect fine-grained meaning differences in human-translated texts, as well as critical errors due to local mistranslations in machine-translated texts.

DETECTING FINE-GRAINED SEMANTIC DIVERGENCES TO IMPROVE TRANSLATION UNDERSTANDING ACROSS LANGUAGES

by

Eleftheria Briakou

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
2023

Advisory Committee:
Professor Marine Carpuat, Chair
Professor Philip Resnik, Dean's Representative
Professor Hal Daumé III, Member
Professor Leo Zhicheng Liu, Member
Professor Luke Zettlemoyer, Member

Acknowledgments

I would like to express my deepest appreciation to all the people who have supported me throughout my journey in completing this thesis.

First, I would like to extend my gratitude to my supervisor, Marine Carpuat. Her guidance, expertise, and feedback have been instrumental in shaping the direction and content of this thesis, as well as shaping my own growth and development as a researcher. I would also like to acknowledge the members of my thesis committee, Luke Zettlemoyer, Hal Daumé III, Philip Resnik, and Leo Zchicheng Liu, for their valuable input and thought-provoking discussions.

Throughout my Ph.D. journey, I was fortunate to have worked with many wonderful researchers. Many thanks to my industrial collaborators and mentors Joel Tetreault, Marjan Ghazvininejad, Luke Zettlemoyer, George Foster, and Colin Cherry for the internship experiences and fruitful collaborations.

My deepest gratitude is reserved for the remarkable group of my fellow labmates in CLIP who have fostered a welcoming and supportive environment. Special thanks to my dearest labmates and friends, Sweta Agrawal and Pranav Goel, who have shared the entire Ph.D. journey with me and have steadfastly stood by my side throughout. I would also like to extend many thanks to Yoo-Yeon Sung, Calvin Bao, Alexander Miserlis Hoyle, Navita Goyal, and all CLIP labmates who joined me later in my Ph.D. journey.

To my friends who have made my transition to the United States as smooth and as enjoyable as it could be: Ioanna Karantaidou, Leda Apergi, Leonidas Tsepenekas, Leonidas Lampropoulos, Kostas Zampogiannis, Christos Papageorgakis, Panagiotis Dimitrellos, Christos Zafeiriadis, Danae Mitka, and Sotiria Stathopoulou. I would like to give a special shoutout to my dearest roommates, Ioanna Galanis, José Luis Villaescusa Nadal, and Stefano Antonini, with whom we navigated the challenging years of the COVID era. Many thanks

to my long-term friends, who have been a constant source of support and encouragement over the past decade: Emmanouela Kapetanaki, Foteini Barka, Christiana Tsiouri, Eleni Zacharopoulou, and Vasiliki Kontou.

Last but not least, I want to thank Ahmet Ozan Sivri, my siblings, Eirini-Niki and Spyros, and my parents for their invaluable support and encouragement throughout this journey. I am truly grateful for your presence in my life and for your belief in me.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Comparing and contrasting the meaning of text conveyed in different languages is a fundamental task in Natural Language Processing (NLP). For instance, detecting the commonalities and divergences between sentences drawn from English and French Wikipedia articles about the same topic can help analyze language bias (Bao et al., 2012; Massa and Scrinzi, 2012), or mitigate differences in coverage and usage across languages (Yeung et al., 2011; Wulczyn et al., 2016; Lemmerich et al., 2019; Johnson and Lescak, 2022). This requires computational approaches that can capture not only coarse content mismatches across languages but also fine-grained differences in sentences that largely convey the same content—we term such mismatches in meaning as *fine-grained cross-lingual semantic divergences*. Such divergences abound in *parallel texts*—texts in two different languages that are perceived as exact translations of each other. As we will see in Section §2, fine-meaning mismatches can be found in parallel corpora that are created in different ways. For instance, in OpenSubtitles—a corpus consisting of automatically aligned movie subtitles from different languages—the English sentence *"How much do you get paid?"* is paired with the divergent sentence *"T'es payé combien de l'heure?"* ("How much do you get paid per hour?") in French. Additionally, nuanced meaning differences can be introduced by human translators when they select between near synonyms (Hirst, 1995). For instance, the English Wikipedia article of Erotokritos states *"Erotokritos is a romance"*, while the corresponding Greek article says "Ο Ερωτόκριτος είναι ένα μυθιστόρημα" ("Erotokritos is a novel"). Thus, the English article is more specific than the Greek article. Regardless

of why such subtle divergences exist in parallel texts, we hypothesize that they matter not only for humans' understanding of the differences in content across languages, but also for machine translation—as they yield challenging training samples and for humans—who might benefit from a nuanced understanding of the source. Thus, in this thesis, we argue that quantifying fine-grained divergences is crucial to **improve both *machine* and *human* translation understanding across languages**.

In Chapter 3, we start our exploration by establishing that fine-grained divergences are frequent in parallel texts. We analyze data from WikiMatrix—a standard resource for training MT systems—via collecting human assessment of semantic divergence in English and French. To that end, we introduce an annotation protocol that asks for human rationales and establish that divergences can be reliably annotated under our protocol. Our annotation findings show that fine-grained divergences are frequent in parallel texts: 40% of WikiMatrix samples are judged as containing fine-meaning mismatches by human annotators. Additionally, we show that manual annotation is expensive and hard to scale, which motivates computational approaches to quantifying divergences at scale. Drawing on our annotation findings, we introduce a contrastive loss designed to make a multilingual language model sensitive to subtle cross-lingual differences between linguistically motivated synthetic samples. After establishing that fine-grained divergences can be detected accurately at scale, our next work explores their impact on Neural Machine Translation (NMT).

In Chapter 4, we contribute an analysis of how fine-grained divergences in training data impact NMT quality and confidence. We provide empirical evidence that these imperfect training references: hurt translation quality (as measured by automatic metrics) once they overwhelm equivalents, output degenerated text more frequently, and increase the uncertainty of models' predictions. Drawing on those findings, in Chapter 5, we seek ways to mitigate the negative impact of fine-grained divergences on NMT and explore two orthogonal strategies. Our first strategy intervenes in the NMT training assumption of translation

equivalence in parallel texts and aims to model divergences explicitly. Our second strategy proposes an orthogonal mitigation direction: instead of altering NMT training to model divergences closely, we aim to automatically re-write divergent samples to yield more equivalent translations. We introduce different models and algorithms along those orthogonal intervention strategies and show that by considering semantic divergences, we can improve machine translation across a wide range of low, medium, and high resource data regimes. After exploring how detecting semantic divergences helps us improve machine translation understanding, our final work contributes ways of assisting *humans* in understanding and detecting translation differences.

In Chapter 6, we explore ways of explaining fine-grained divergences to humans. We start our exploration by surveying relevant literature on explaining NLP predictions. Then, we introduce an approach that explains the prediction of our divergent models by producing contrastive phrasal highlights, which can be color-coded to convey mappings between a source text and its translation intuitively. We show that contrastive phrasal highlights match human-provided rationales of semantic divergences better than standard highlighting approaches that ignore the input's structure. Furthermore, through a series of human-centered evaluations, we provide evidence that our highlights assist bilingual speakers in *annotating* fine-grained meaning divergences in human-translated pairs reliably and, therefore, ease the need for human rationales. Additionally, we show that contrastive phrasal highlights help bilingual speakers in *detecting* critical errors due to local mistranslations in machine-translated texts more accurately.

Finally, in Chapter 7, we conclude this dissertation thesis by summarizing our findings and contributions and highlighting avenues for future work.

Chapter 2: Background

This work draws on many different areas in machine translation and semantics. We first discuss the fundamental concepts and techniques behind acquiring parallel texts, along with recent efforts to audit them, in Section 2.1. We then discuss the use of parallel texts as critical data resources in Machine Translation and Multilingual NLP research in Section 2.2. Finally, we discuss the literature related to the detection of cross-sentence meaning mismatches within and across languages in Section 2.3.

2.1 Acquiring and Auditing Parallel Texts

Parallel texts have been the object of systematic analysis by translation scholars in the field of Translation Studies (Floros, 2004) and have also been explored as a pedagogical tool in foreign language acquisition in Social and Education Studies (Bluemel, 2014; Abdallah, 2021; Awad et al., 2021). Apart from their direct use in the above fields, parallel texts emerged as a valuable resource in Computational Linguistics, enabling the development of NLP technology that breaks language barriers. The definition of parallel texts has long been debated in translation studies, with some of them identifying them as original texts in different languages that are assigned an interlingual dimension, while others treat them as translations. In this thesis, we use the term parallel text, also known as bitext, as used in Computational Linguistics, i.e., *an original text paired with its translation*. Thus, a parallel corpus consists of two texts: the source text and its translation. One of the major bottlenecks

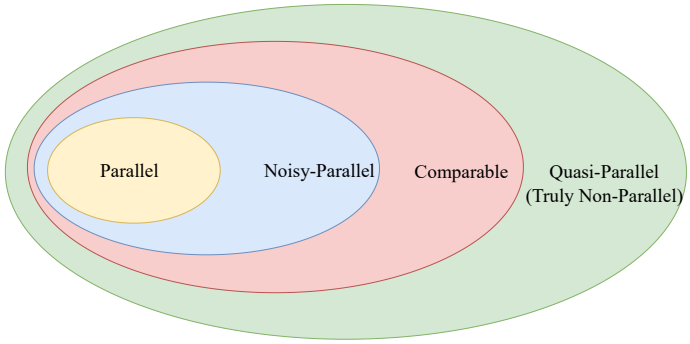


Figure 2.1: Venn diagram illustrating the relation between (non)-parallel corpora.

in acquiring large amounts of parallel texts has been the human labor-intensive task of translation which places responsibility on multilingual humans fluent in multiple languages and cultures. Crucially, reliance on human translations poses a severe challenge for under-resourced languages for which multilingual humans might be limited or even not available. The same problem also arises in better-resourced languages when translation is required in narrow, specialized domains, such as medicine, and information technology, among others. As a result, it is crucial that technological solutions to acquiring parallel texts are found to compensate for the shortage of linguistic resources for under-resourced languages and specialized domains. Non-parallel corpora—bodies of texts in two different languages—stand as a fundamental resource in acquiring parallel texts, constituting a long-standing natural language processing research problem. In this section, we begin our discussion with an overview of the degrees of parallelism in non-parallel corpora in §2.1.1; we then summarize the most fundamental computational approaches to parallel text acquisition in §2.1.2 and finally discuss efforts to audit extracted parallel texts in §2.1.3.

2.1.1 Degrees of Non-Parallel Corpora

The degree of parallelism across (non)-parallel corpora might vary depending on the underlying curation processes. *Was each body of text written independently in each lan-*

guage? Was one of them generated as a translation of the other? If so, what was the granularity at which translation took place, e.g., article vs. document level? Although answering those questions precisely is hard, given the explosion of undocumented online content worldwide, they provide a foundation for framing the discussion around the different degrees of parallelism we can encounter amongst non-parallel corpora. Formally, bilingual corpora can be classified into four main categories along the dimension of their parallelism (Fung and Cheung, 2004a,b; Wu and Fung, 2005):

Parallel

DEFINITION: A *sentence-aligned* corpus containing translations of the same document.

EXAMPLE: The Hong Kong Laws Corpus is a parallel corpus with manually aligned sentences (Fung and Cheung, 2004a).

Noisy Parallel

DEFINITION: A non-sentenced-aligned corpus that nevertheless contains mostly bilingual translations of the same document.

EXAMPLE: The Hong Kong News corpus contains documents that are rough translations of each other, focused on the same thematic topics, with some insertions and deletions of paragraphs (Fung and Cheung, 2004b).

Comparable

DEFINITION: A *non-sentence-aligned* corpus that contains documents that are topic-aligned but are not translations of each other.

EXAMPLE: Newspaper articles from two sources in different languages within the same window of published dates, can constitute a comparable corpus.

Quasi-Parallel (Truly Non-Parallel)

DEFINITION: A corpus that contains non-aligned and non-translated bilingual documents that could either be on the same topic (in-topic) or not (off-topic).

EXAMPLE: TDT3 Corpus is a good source of truly non-parallel and quasi-comparable corpus. It contains transcriptions of various news stories from radio broadcasting or TV news report from 1998-2000 in English and Chinese.

2.1.2 Acquiring Parallel Text from Non-parallel Corpora

Acquiring parallel texts from non-parallel corpora, such as web crawls, is a long-standing NLP problem (Resnik, 1999). In principle, parallel text acquisition is operationalized through a series of pipelined subtasks, ranging from heavily engineering processes of web crawling to addressing fundamental cross-lingual NLP subtasks. Typically, the main steps involved in parallel text acquisition from non-parallel corpora are web crawling, document alignment, sentence alignment, and filtering. Depending on the nature of the web-crawled data, only some of the steps take place in delivering the final "clean" parallel corpus, as seen in Table 2.1. Below we discuss how each of the steps involved in parallel texts acquisition is addressed in NLP research.

Document Alignment In the task of document alignment, the goal is to start from a given collection of documents and identify pairs of documents in two languages such that one document is the translation of the other. Early practices include computing edit-distance scores between linearized documents using information about document structure (Resnik and Smith, 2003) or hand-crafted rules that rely on metadata such as publication dates (Munteanu and Marcu, 2005, 2006). Subsequent methods rely, instead, mostly on the textual contents of the given documents. For instance, Shi et al. (2006) use probabilities of a probabilistic tree alignment model, while Uszkoreit et al. (2010) compute cosine distance of inverse document frequency weighted n-gram vectors. In the shared task for bilingual document

Parallel Corpus	Web Crawling	Document Alignment	Sentence Alignment	Bitext Filtering
EuroParl	✓	✓	✓	✗
ParaCrawl	✓	✓	✓	✓
TED	✓	✗	✗	✗
OpenSubtitles	✗	✗	✓	✓
WikiMatrix	✓	✗	✓	✓
CCMatrix	✓	✗	✓	✓

Table 2.1: Steps involved in acquisition of large-scale parallel OPUS corpora (Tiedemann, 2012b).

alignment (Buck and Koehn, 2016a), many participants use techniques based on machine translation models, bag-of-words lexical translation probabilities, or document features for scoring candidate document pairs. Concretely, methods based on machine translation are used used to map the collection of French documents into English and then cast the problem as a monolingual scoring task. Submitted systems suggest computing coverage scores of a document-pair based on the lexical overlap of phrases (Gomes and Pereira Lopes, 2016), n-gram matches (Dara and Lin, 2016; Shchukin et al., 2016), and word occurrence probabilities (Le et al., 2016). Other approaches rely on computing cosine similarity scores between term frequency-inverse document frequency (td/idf) weighted vectors extracted by collecting n-grams on documents that are mapped to the same language via machine translation (Buck and Koehn, 2016b) or word translation lexicons (Azpeitia and Etchegoyhen, 2016; Medved' et al., 2016; Jakubina and Langlais, 2016; Mahata et al., 2016). Finally, some participants propose document alignment based on feature extraction, such as links to documents in the same web domain, similarity of link URLs, HTML tags, matched named entities (Papavassiliou et al., 2016; Esplà-Gomis, 2009; Esplà-Gomis et al., 2016; Lohar et al., 2016).

Sentence Alignment In the task of sentence alignment, the goal is to start from aligned documents—produced as (human) translations of each other *or* automatically extracted

using document alignment approaches—and find matches of sentences that are translations of each other. Early methods are based on sentence length, measured in words or characters in each sentence (Brown et al., 1991; Gale and Church, 1993). In contrast, later approaches are based on optimizing word-translation probabilities via integrating lexical information from bilingual word dictionaries that are either provided (Chen, 1993; Moore, 2002; Varga et al., 2007), or extracted in an unsupervised fashion (Braune and Fraser, 2010). Later work use machine translation to map both documents in the same language and introduce scoring methods based on their lexical overlap (Sennrich and Volk, 2010; Gomes and Lopes, 2016). Recent advances in multilingual representation learning (Artetxe and Schwenk, 2019; Liu et al., 2020) enable the extraction of parallel texts across multiple languages based on normalized cosine distance metrics Thompson and Koehn (2019); Schwenk et al. (2021a,b).

Bitext Filtering In the task of parallel text filtering, also known as bitext filtering, to goal is to start from a set of web-crawled *noisy parallel* corpus and filter it to a smaller size of high-quality sentence pairs, to deliver a final "clean" corpus, i.e., a *parallel* corpus under the definitions provided in §2.1.1. This task has received significant NLP interest which led to a dedicated shared task of "Parallel Corpus Filtering" in WMT (Annual Conference of Machine Translation) since 2018 (Koehn et al., 2018, 2019, 2020). Past submissions to the shared task employ a diverse set of approaches covering simple pre-filtering rules based on language identifiers and sentence features (Rossenbach et al., 2018; Lo et al., 2018; Ash et al., 2018) learning to weight scoring functions based on language models, extracting features from neural translation models and lexical translation probabilities (Sánchez-Cartagena et al., 2018), combining pre-trained embeddings (Papavassiliou et al., 2018), and dual-cross entropy (Chaudhary et al., 2019).

2.1.3 Auditing Parallel Texts Extracted from (Non)-Parallel Corpora

Although extracting parallel texts from non-parallel corpora appears as a tempting solution to compensate for the shortage of resources in specific languages and domains, the resulting sentence pairs might not be as “parallel" as we usually assume they are. Despite the extensive literature on parallel text acquisition, the quality of the resulting pairs is mostly implicitly documented through their use in machine translation and multilingual NLP research, as we extensively discuss in Section §2.2. Few recent works constitute an exception to this extrinsic evaluation framing and follow an intrinsic evaluation instead, based on manual auditing of samples of parallel texts. We summarize those efforts in Table 2.2, present examples of parallel texts identified as problematic in Table 2.3, and discuss them in detail below.

CITATION	LANGUAGE-PAIRS	PARALLEL CORPUS
Khayrallah and Koehn (2018)	English—German	ParaCrawl v.1
Kreutzer et al. (2022b)	English—{Uyghur, Mirandese, Tajik, Nepali, Georgian, Lombard, Ido, Javanese, Wu Chinese, Breton, Bavarian, Kazakh, Swahili, Low German, Belarusian, Hindi, Korean, Ukrainian, Italian, English simlified}	WikiMatrix
	English—{Somali, Pashto, Burmese, Khmer, Nepali, Swahili, Sinhala, Norwegian Nynorsk, Basque, Galician, Russian, Bulgarian, Catalan, Greek, Polish, Dutch, Portuguese, Italian, Spanish, German, French}	ParaCrawl v7.1
Vyas et al. (2018a)	English—French	OpenSubtitles, CommonCrawl
Zhai et al. (2018)	English—{French, Chinese}	TED Talks

Table 2.2: Parallel corpora along with the complete set of language pairs that have been (sampled and) manually audited in prior work.

Auditing for Noise

Khayrallah and Koehn (2018) audit for several types of noise in 200 English-German parallel texts sampled from the ParaCrawl corpus (Esplà et al., 2019; Bañón et al., 2020). The sample is classified into one of five error categories: misaligned sentences (i.e., German

and English sentences that are not translations of each other), third language (i.e., wrong language on either side), untranslated (i.e., empty sequences on either side), copy (i.e., the same sentence appears in both sides), short-segments, and non-linguistic content (i.e., sentences that contain junk sequences).

→ **Parallel Texts** The ParaCrawl v.1 corpus contains sentence pairs extracted from the Internet for languages used in the European Union and thus is an example of parallel text acquisition from *quasi-parallel* corpora. First, parallel documents are discovered via downloading HTML candidate documents from the Internal that are further aligned based on the cosine distance of tf/idf weighted document vectors. Then, parallel texts are extracted by aligning the segments in each of the document pairs identified that are further refined by filtering. The mined pairs do not necessarily constitute human-translated text and are expected to be extremely noisy.

→ **Findings** Their audit shows that parallel texts in ParaCrawl v.1 are noisy 77% of the times. The most frequent source of noise is that of misaligned sentences (41%), followed by a noise that reflects wrong language content on any of the sides (23%). Untranslated sentence pairs correspond to 4% of the annotated sample, while 2% of sentence pairs consist of random byte sequences, only HTML markup, or Javascript. Finally, a number of sentence pairs have very short German or English sentences, containing at most 2 tokens (1%) or 5 tokens (5%).

Kreutzer et al. (2022b) manually audit 100 samples from several multilingual datasets in 205 language-specific corpora that result from automatic curation pipelines, including parallel texts from 20 language-pairs in WikiMatrix (Schwenk et al., 2021a) and 21 language-pairs from ParaCrawl v7.1 (Bañón et al., 2020; Esplà et al., 2019). Their primary goal is to audit

the quality of multilingual datasets to uncover whether systematic issues arise with a focus on the lowest-resourced and under-evaluated languages. To that end, they develop a simple error taxonomy based on three main classes: Incorrect Translation, Wrong Language, and Non-Linguistic Content, while for correct sentences, they further mark single words or phrases and boilerplate (i.e., low-quality) contents.

→ **Parallel Texts** The WikiMatrix corpus (Schwenk et al., 2021a) contains sentence pairs extracted from the large collection of Wikipedia pages in two different languages that are mainly considered topic-aligned; it is, thus, an example of parallel texts acquisition from *comparable* corpora. Extraction of parallel texts is based on a distance-based mining approach that uses cross-lingual similarity in a language-agnostic sentence embedding space. Although a Wikipedia page in one language could result from (human) translation of another Wikipedia page, WikiMatrix is based on a global mining approach that considers the whole Wikipedia for each language instead of performing local mining within translated articles. As a result, it is unclear whether parallel texts in WikiMatrix are produced as translations of each other in the first place. The ParaCrawl v7.1 corpus is curated following the same pipeline as in ParaCrawl v.0 (see ¶Auditing for Noise), including more languages and improved technologies for implementing each pipeline component. As a result, subsequent versions of ParaCrawl are expected to be less noisy than their earlier counterparts.

→ **Findings** Their audit reveals systematic issues for several low-resource languages: parallel texts in WikiMatrix correspond to incorrect translations more than 80% of the times for 10 out of the 20 audited language-pairs; while incorrect parallel texts overwhelm the ParaCrawl v7.1 corpus $> 50\%$ for 5 out of the 21 audited pairs. In principle, incorrect translations are the most frequent source of low-quality reported in most pairs, followed by wrong language and non-linguistic content.

Auditing for Translation Equivalence

Vyas et al. (2018a) audit 300 English-French sentence pairs drawn from two well-studied resources of parallel texts, i.e., the OpenSubtitles (Lison and Tiedemann, 2016) and the Common Crawl corpus. They ask bilingual annotators to provide their judgment on whether they agree or disagree with the statement "the French and English text conveys the same information."

→ **Parallel Texts** The OpenSubtitles corpus contains sentence pairs extracted from a *noisy parallel* corpus consisting of translations of movie subtitles. Although it contains human translations, mined sentence pairs are not always expected to be completely parallel given the many constraints imposed on translations: texts should fit on a screen and be synchronized with a movie. Furthermore, using more informal registers often requires frequent non-literal translations of figurative language. The Common Crawl corpus contains sentence pairs extracted from web crawls that fall under the *quasi-parallel* corpus definition. Parallel texts are automatically mined from sentence-aligned parallel documents on the Internet. Parallel documents are discovered using, e.g., URL containing language code patterns, and sentences are automatically aligned after structural cleaning of HTML. The mined sentence pairs do not constitute human-translated text of each other, and the sentence pairs are expected to be extremely noisy.

→ **Findings** Their audit shows that texts perceived as parallel contained meaning differences 43.6% and 38.4% of the times in OpenSubtitles and Common Crawl, respectively. The meaning mismatches found in the studied parallel texts ranged from completely unrelated texts to ones that contain more subtle differences, e.g., "what does it mean" in English vs. "what are the advantages" in French.

Auditing for Translation Processes

Zhai et al. (2018) audit human translations in English to French and Chinese parallel texts of transcribed and translated talks, including in the corpus of TED talks (Cettolo et al., 2012). Their audit focuses on identifying translation processes at a phrasal level apart from literal translations. They propose a hierarchy of relations that classifies a translation unit (e.g., phrase) as belonging to one of four categories: *literal*, *equivalence*, *non-literal*, and *unaligned*. They argue that specific types of non-literal translations could result in semantic shifts. For instance, in the case of *particularization*, the translation is more precise or presents a more concrete sense than that of the source, e.g., “language loss" in English can be translated in “l’extinction du langage" (i.e., the extinction of language) in French.

→ **Parallel Texts** The TED Talks corpus is released for the evaluation campaign of the International Workshop on Spoken Language Translation (IWSLT) (Cettolo et al., 2013, 2014). The source language (the original language in which the speakers expressed themselves) is English. At the same time, the translation of subtitles for TED Talks is controlled by volunteers and language coordinators per language. This process generally ensures good quality translations; thus, the corpus is considered *parallel*.

→ **Findings** They find that non-literal translations are more frequent in translating to Chinese than French: 47% vs. 32% of translated tokens are marked as belonging to a non-literal translation process. Furthermore, 8% of French translated tokens (and 9% of Chinese) correspond to translation techniques that introduce semantic shifts, while 1.5% of French tokens (and 5% for Chinese) are removed deliberately in translation. The latter denotes a clear difference between the *simplification* translation process that takes place in Chinese vs. French translations, as Chinese translations often convey only the most important information and leave some other content phrases non-translated.

Incorrect translation, but both correct languages (Kreutzer et al., 2022b)			
ENGLISH	A map of the arrondissements of Paris	KONGO	Paris kele mbanza ya kimfumu ya Fwalansa.
ENGLISH	Ask a question	TURKISH	Soru sor Kullanıma göre seçim
Source OR target wrong language, but both still linguistic content (Kreutzer et al., 2022b)			
ENGLISH	The ISO3 language code is zho	DIMILI	Táim eadra bracach mar bhionns na frogannaidhe.
ENGLISH	Der Werwolf — sprach der gute Mann,	GERMAN	des Weswolfs, Genitiv sodann,
Not a language: at least one of source and target are not linguistic content (Kreutzer et al., 2022b)			
ENGLISH	EntryScan 4 _	TSWANA	TSA PM704 _
ENGLISH	organic peanut butter	KURDISH	? ? ? ? ? ? ?
Non-linguistic content (Khayrallah and Koehn, 2018)			
ENGLISH	Anonymous 2 2010-03-24 at 20:55 314 Comments	GERMAN	Anonym 2 24.03.2010 um 20:55 314 Kom- mentare
ENGLISH	< < first < prev. page 3 next last > >	GERMAN	< < zuerst & lt; vorh. Seite 3 nächste letzte
Meaning differences (Vyas et al., 2018a)			
ENGLISH	what does it mean when food iss “low in ash" or “low in magnesium"?	FRENCH	quels sont les avantages dune nourriture “réduite en cendres” et “faible en magnésiumm” ?
ENGLISH	rabbit? if i told you it was a chicken, you wouldn’t know the difference	FRENCH	vous croirez manger du poulet.
Translation Processes (Zhai et al., 2018)			
ENGLISH	and you’ll suddenly discover what it would be like	FRENCH	et vous découvrirez ce que ce serait
ENGLISH	this is a people who cognitively do not distinguish	FRENCH	c’est un peuple dont l’état des connaissances ne permet pas de faire la distinction

Table 2.3: Examples of manually audited parallel texts that found to be “non-parallel".

2.2 Using Parallel Corpora for Multilingual NLP

Parallel texts have been a critical resource for training NLP models by providing either direct or distant supervision. At the core of their use lies the assumption that the two pieces of text represent a relation of exact meaning equivalence. As we discuss below, this assumption is either modeled explicitly, i.e., in supervised Neural Machine Translation (Section 2.2.1) or is implicitly made when parallel texts are used as a form of distant supervision in several other multilingual NLP applications (Section 2.2.2).

2.2.1 Parallel Texts in Neural Machine Translation

Machine Translation (MT) is the task of automatically translating a piece of text from one language to another. Among the many different approaches to MT, Neural Machine Translation (NMT) has emerged as the most promising one, leading to rapid adoption in deployments, e.g., Google (Wu et al., 2016), Systran (Crego et al., 2016), and WIPO (Junczys-Dowmunt et al., 2016). However, at the core of NMT's success lies the availability of large-scale training data, i.e., parallel texts (Koehn and Knowles, 2017). The latter provides direct supervision signals for NMT models that are typically trained to maximize the log-likelihood of the training data, $\mathcal{D} \equiv \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$, where $(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})$ is the n-th parallel text. Concretely, the core assumption made when using parallel text to train supervised NMT models is that the sentences $\mathbf{x}^{(n)}$ and $\mathbf{y}^{(n)}$ are exact translations of each other. This assumption is operationalized via training a single neural model with parameters θ to maximize the token-level cross-entropy loss:

$$\mathcal{J}(\theta) = \sum_{n=1}^{N} \sum_{t=1}^{T} \log p(y_t^{(n)} \mid \mathbf{y}_{<t}^{(n)}, \mathbf{x}^{(n)}; \theta) \tag{2.1}$$

Yet, as discussed in Section 2.1.3 parallel texts might violate the translation equivalence

assumption for different reasons and to different extents (e.g., translation processes vs. noise). Khayrallah and Koehn (2018) focus on the extreme case where the translation equivalence assumption is violated due to *noise* in the training data and explore how various types of noise impact on NMT. Concretely, they contribute an empirical analysis based on five types of artificial noise and analyze how they degrade performance in neural and statistical machine translation systems. They find that neural models are generally more harmed by noise than statistical models and that untranslated training instances cause NMT models to copy the input sentence at inference time. Additionally, Ott et al. (2018) argue that miscalibration of NMT models can be attributed to the "extrinsic" uncertainty of the noisy, untranslated references found in the training data. The above findings motivated a shared task dedicated to filtering noisy samples from web-crawled data at WMT, since 2018 (Koehn et al., 2018, 2019, 2020).

Even after noise filtering, a wide range of the remaining samples contains meaning mismatches. Recent work has effectively focused on coarse-grained meaning differences in parallel texts: Vyas et al. (2018a) work on subtitles and Common Crawl corpora where sentence alignment errors abound. They provide empirical evidence that coarse semantic divergences matter for NMT and show that filtering out divergent examples helps speed up the convergence of neural NMT training without loss in translation quality. In a similar spirit, Pham et al. (2018) focus on automatically fixing a specific divergent type that is frequent in translations of subtitles corpora, i.e., divergences where content is appended to one side of a translation pair. They show that fixing those divergences yields alternative parallel texts that provide a more reliable training signal for NMT than training on the originals or even a filtered version of the corpus that disregards them.

2.2.2 Parallel Texts in Multilingual NLP

Beyond their use in MT, parallel texts have been long used in several other areas of multilingual NLP with different motivations ranging from—extracting cross-language information, e.g., word alignments or dictionaries to—providing seed instances for the generation of synthetic training samples for other multilingual or cross-lingual NLP tasks, e.g., automatic post-editing. Across applications, and *similar to* MT, parallel texts are again treated as exact translations; however, *unlike* MT, parallel texts are mostly used as an indirect form of supervision. Below, we discuss some popular examples of using parallel text in multilingual NLP research beyond MT.

Alignment and Tagging for Linguistic Knowledge Extraction A wide range of NLP research is based on word alignment and tagging to induce cross-lingual or multilingual annotations at a sub-sentential level. At the core of most approaches lies the fundamental task of *bitext word alignment* i.e., finding links between words given pairs of translated sentences (Tiedemann, 2011). Word aligners traditionally require parallel texts for training. Statistical word aligners such as the IBM models (Brown et al., 1993) and their implementations Giza++ (Och and Ney, 2003), fast-align (Dyer et al., 2013) define a probabilistic translation model and estimate its parameters based on the observed parallel texts using the expectation-maximization algorithm. Despite statistical aligners that model the translation equivalence assumption within parallel texts closely, more recent aligners rely on parallel texts implicitly. For instance, Peter et al. (2017); Zenkel et al. (2019) extract word alignments based on NMT attention matrices while Garg et al. (2019) pursue a multitask approach for alignment and translation.

The use of word alignments has been explored on a series of downstream NLP applications. Shi et al. (2021) extract features based on global statistics of word alignments extracted

from parallel texts to induce bilingual word lexicons. They provide empirical evidence that the assumption of equivalence matters by showing that the lexicon induction performance correlates well with the quality of parallel texts. Another thread of works combines word alignments and parallel texts to perform cross-lingual annotation projection in an attempt to relieve the effort involved in creating annotations for new languages. Such approaches create new monolingual resources by transferring annotations from one language (typically English or other high-resource languages) onto resource-scarce languages, assuming that parallel texts are exact translations of each other. Examples include projections of parts of speech tags (Yarowsky and Ngai, 2001; Hwa et al., 2005), chunks (Yarowsky et al., 2001), syntactic dependencies (Hwa et al., 2002), word senses (Diab and Resnik, 2002; Bentivogli and Pianta, 2005; Ettinger et al., 2016), semantic roles (Padó and Lapata, 2009), among others. Although the impact of non-parallelism in parallel texts used for cross-annotation tasks has not been extensively quantified, it is implicitly taken into account via filtering the resulting annotations to reduce the impact of alignment noise.

Multilingual Representation Learning Recent works explore the impact of incorporating parallel texts in multilingual language (Devlin et al., 2019; Conneau et al., 2020) and sentence modeling (Artetxe and Schwenk, 2019). Such models are typically used to extract universal representations across languages that are further used to improve the performance of cross-lingual natural language understanding (NLU) tasks via fine-tuning on dedicated tasks. At pre-training time, the models optimize a combination of several monolingual and cross-lingual language model objectives. Monolingual objectives aim at reconstructing input texts via defining several noising functions. In contrast, cross-lingual objectives typically take the form of translation modeling (i.e., Equation 2.1) and, thus, use parallel text as a direct supervision. Although recent works provide evidence that parallel texts benefit cross-lingual transfer in pre-trained language models (Conneau and Lample, 2019; Hu et al.,

2021; Ouyang et al., 2021; Luo et al., 2021; Kale et al., 2021) as measured by downstream performance gains on NLU tasks, it is not clear whether and how the non-parallelism in parallel texts has a measurable impact on their downstream performance.

Synthetic Data Generation Parallel texts have been used as seed translations for generating synthetic samples to train a wide range of generation and sequence tagging tasks. For instance, *word-level quality estimation* (word-QE)—the task of tagging an MT output with labels indicative of good vs. bad translations and *automatic post-editing* (APE)—the task of editing an MT output to match better the semantics of the source, typically require supervised data for training. Yet, due to the scarcity of gold-standard data for such tasks, a standard approach is that of augmenting them with synthetic samples that are automatically generated (Chatterjee et al., 2018, 2019, 2020). Concretely, synthetic supervision is based on using parallel texts as exact translations and extracting silver bullet annotations using external MT models and word aligners (Negri et al., 2018). Finally, parallel texts have also been explored as an indirect form of supervision in multi-task or zero-shot settings in monolingual text generation tasks, such as style transfer (Niu et al., 2018; Briakou et al., 2021) and paraphrase generation (Thompson and Post, 2020), where their primary motivation is that of encouraging cross-language transfer learning.

2.3 Semantic Divergences Within and Across Languages

2.3.1 Defining Semantic Divergences

Studying semantic divergences requires a definition of the concept of *semantic equivalence*, i.e., two sentences are semantic equivalent if they represent the same event or fact without additional explanation of the context in which these two sentences occur. In the context of Natural Language Processing, much work views translation as a cross-lingual

mapping that adheres to the properties of semantic equivalence. Yet, when humans translate, they resort to processes that might deliberately introduce *fine-grained semantic divergences*, i.e., meaning differences to account for context, such as who the intended target audience is or what the intended use of the translation is. Drawing on those insights, in this thesis, we depart from the view that translations are either correct or noisy based on their semantic equivalence and rather view translation equivalence as a continuum that allows us to account for contextual cues and fine-grained semantic divergences as a tool that helps us characterize this continuum better.

An example of a (human) translation that falls into this continuum is the English translation of “The Maple Leaf Forever is a national anthem" drawn from Wikipedia, which maps to “Maple Leaf Forever est un chant patriotique" (i.e., The Maple Leaf Forever is a patriotic song) in French. Unlike prior work that focuses on coarse content mismatches across languages, which are primarily seen as translation errors, for instance, in the context of cross-lingual similarity or machine translation quality estimation settings, such translations diverge in meaning in more *nuanced* ways to potentially reflect variations in cultural perspectives and terminology. We thus argue that perceiving them as errors is a restricted perspective. At the same time, those semantic divergences differ from *translation divergences*, which are defined as structural differences in translations that convey the exact same meaning (Dorr, 1994).

In the rest of this chapter, we start by establishing the foundations of how sentence meaning is operationalized in this thesis in Section §2.3.2. Then, we shift our discussion to prior work on detecting cross-lingual semantic divergences in the context of parallel corpora in Section §2.3.3. Finally, we point to a large body of work that looks at the broader notion of semantic equivalence both within and across languages in Section §2.3.4.

2.3.2 Operationalizing Semantic Divergences

To understand how we computationally study the concept of semantic equivalence, we start by defining the meaning of a word as its word senses (Murphy, 2010). Based on this definition, two words are equivalent in meaning if they share at least one word sense. In the context of NLP, we use computational approaches that represent words as distributional or vector space representations. Such approaches are based on the distributional hypothesis (Harris, 1954), which presumes a correlation between distributional similarity and meaning similarity. The direct implication of this hypothesis is that two words considered semantically similar are expected to occur in similar contexts, which we define as the set of words existing within a window around each occurrence of the target word. Then, we move from the word-level operationalization of meaning to larger units of texts, such as sentences, by following the principle of compositionality, i.e., the meaning of a sentence is determined by the meanings of its constituent words and the rules used to combine them. Again, we represent the meaning of a sentence using an empirical model that uses attention-based neural networks to encode the meaning of a sentence based on its words as a vector in a high-dimensional space.

2.3.3 Detecting Cross-lingual Semantic Divergences

In this thesis, we follow Vyas et al. (2018a) and define cross-lingual semantic divergences as mismatches in meaning between parallel texts (defined in Section §2.1) that are typically expected to be aligned across languages and equivalent in meaning. Semantic divergences differ from translation divergences that reflect different ways of encoding the same information across languages (Dorr, 1994). We note that although translation divergences and semantic divergences might co-occur, one does not entail the other, as presented in the below examples:

TRANSLATION DIVERGENCE		SEMANTIC DIVERGENCE	
ENGLISH	*I am hungry.*	ENGLISH	*Your father is French.*
GERMAN	*Ich habe Hunger*	FRENCH	*Votre père est Francais.*
GLOSS	*I have hunger.*	GLOSS	*Your parent is French.*

The German translation (left-most example) represents the exact same meaning as the one encoded in the English text. Yet, the two sentences reflect structural divergences—the predicate is adjectival (hungry) in English but nominal (Hunger) in German. On the other hand, the French translation (right-most example) reflects a difference in meaning, as the subject (father) in English is translated to its hypernym (père) in French. At the same time, there are no structural differences between the English text and its translation.

As we reviewed in Section 2.1.3, several efforts to audit parallel corpora indicate that cross-lingual semantic divergences might vary in their granularity for reasons that range from—noise in automatic curation of parallel texts to—(human) translation processes. Below we review prior work on detecting divergences of different granularities at a sentence- and phrase-level.

Unsupervised Detection of Sentence-level Divergences Carpuat et al. (2017) introduce the task of cross-lingual semantic divergence detection as a binary prediction task, i.e., equivalence vs. divergence. They hypothesize that equivalent parallel texts have similar length ratios as measured by the number of words and are easier to align. Based on this hypothesis, they propose an SVM detector that uses sentence length and word alignments as features. They acquire semantic supervision for their task by re-purposing related cross-lingual semantic annotations and evaluate it extrinsically by showing that filtering out divergent pairs benefit NMT training.

In their follow-up work, Vyas et al. (2018a) focus on detecting coarse-meaning differences in OpenSubtitles and CommonCrawl corpora that are discussed in Section §2.1.3. To that end, they introduce a neural cross-lingual semantic similarity model that is designed for

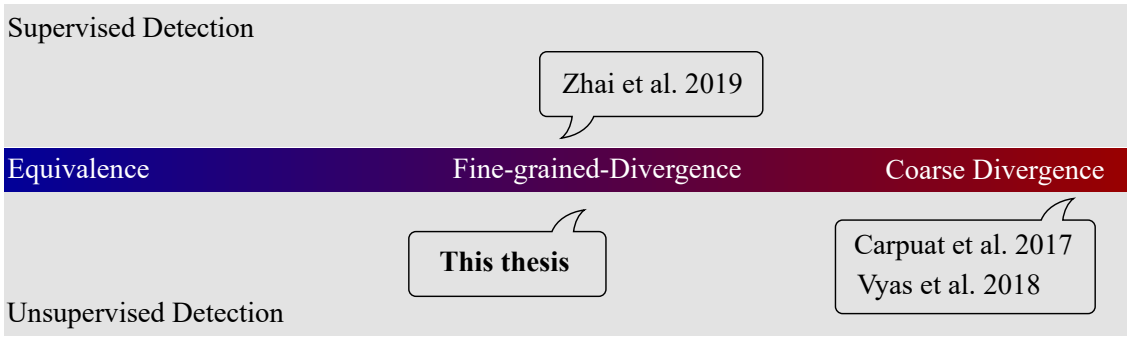
comparing the meaning of sentences in different languages based on cross-sentence word- and span-level similarity comparisons, where a deep convolutional network learns the latter. Cross-language similarity comparisons are enabled by using bilingual word embedding representations that are aligned to lie in the same vector space. Given the lack of supervised data for the task at hand, they devise an automatic way of generating positive (equivalent) and negative (divergent) samples to train their semantic similarity model. First, parallel texts from a parallel corpus are sampled and treated as positive examples. In contrast, candidate negative examples are generated starting from positives by taking the Cartesian product of the two sides of the positive examples. To ensure that the negative set contains hard examples, they perform further filtering so that negative pairs are close to each other in terms of length and word translation coverage under a bilingual dictionary.

On a similar spirit, Pham et al. (2018) propose an unsupervised model for building cross-lingual sentence embeddings are based on modeling word similarity and rely on a neural architecture. Their model computes word-to-word similarities based on contextualized word embeddings that are encoded using bidirectional LSTM neural networks. Since their main focus is to detect divergences in parallel texts due to sentence alignment errors, they train their model to maximize word alignment scores between words in both sentences, using aggregation functions that summarise the alignment scores for each word pair. Similar to Vyas et al. (2018a), the model is trained on both positive and negative samples that are automatically generated via corrupting seed parallel texts.

Supervised Detection of Phrasal Divergences As discussed in Section 2.1.3, semantic divergences can naturally result from different non-literal translation processes. Zhai et al. (2019b) propose an automatic classification of translation processes at the sub-sentential level based on annotated examples collected in their prior work (Zhai et al., 2018). Given a phrasal translation, they frame the task as a multi-class classification problem that predicts

one of the 5 non-literal translation processes: *Equivalence*, *Transposition*, *Generalization*, *Particularization*, *Modulation*, where the latter three translation processes are indicative of semantic divergences. They build a Random Forest statistical classifier based on a rich set of features that leverages syntactic information, semantic similarity scores based on pre-trained word embedding, length ratios, word alignments, and relation links between words extracted from a multilingual knowledge graph.

As reviewed above, prior work addresses cross-lingual semantic divergence detection following two directions: (a) unsupervised detection of coarse meaning differences in parallel texts based on synthetic supervision and (b) supervised detection of phrasal divergences based on gold standard annotations of translation processes. The latter requires an expensive annotation process that does not scale easily. In this thesis, we will address this gap by providing unsupervised detection approaches for detecting divergences of varying granularities, covering both fine- and core-meaning mismatches in parallel texts.



2.3.4 Semantic Divergences Beyond Parallel Texts

This thesis connects to a large body of work that looks at a broader notion of semantic equivalence and has led to defining different framings for capturing cross-sentence semantic relations within and across languages. Although these semantic relation tasks have similar

goals with semantic divergences studied in this thesis, there is an important distinction between them: most cross-lingual semantic relations tasks rely on dedicated datasets that are prepared by translating one side of existing datasets into the language of interest— our goal, instead, is to study divergences in the wild, i.e., in the context of *naturally occurring* parallel texts. We discuss related semantic relation framings along with computational approaches proposed to solve such tasks below.

Semantic Relations Framings Several semantic relation framings have been proposed by prior work aiming at capturing either the different degrees or the different nature of semantic similarity between texts. These frameworks have been used to study semantic relations beyond equivalence between texts within and across languages, where the latter contrasts the meaning of texts in two different languages. We discuss the main questions each of them addresses below and summarize their language coverage in Table 2.4. The task of *paraphrase identification* aims to identify whether two pieces of text have the same meaning or not (Dolan et al., 2004). Although the logical definition of paraphrases requires strict equivalence, Bhagat and Hovy (2013) argue that relaxing this requirement and accepting a broader notion of equivalence, allowing for more examples of "quasi-paraphrase", is more in line with the way paraphrases are understood in linguistics literature. Concretely, they define quasi-paraphrases as "sentences or phrases that convey approximately the same meaning using different words" and discover a set of 25 classes of paraphrasing phenomena via analyzing paraphrases in different corpora. Under their definition, quasi-paraphrases can reflect fine-grained semantic divergences due to, e.g., hypernym/hyponym substitutions ("Pat is flying in this *weekend*" $\Longleftrightarrow$ "Pat is flying in the *Saturday*"), semantic implication ("Google *is in talks to buy* Youtube" $\Longleftrightarrow$ "Google *bought* Youtube"), external knowledge ("The *government* declared victory in Iraq" $\Longleftrightarrow$ "*Bush* declared victory in Iraq"). Furthermore, semantic divergences have also been recorded by Pavlick et al. (2015) who

analyzed paraphrases in ParaPhrase DataBase (PPDB) (Ganitkevitch et al., 2013) a key data resource for this task, comprising millions of automatically extracted paraphrases in multiple languages through bilingual pivoting (Bannard and Callison-Burch, 2005). Concretely, they show that divergent pairs might represent a variety of relations, including directed entailment (*little girl*/*girl*) and exclusion (*nobody*/*someone*). They automatically assign semantic relations to entries in PPDB using the neural logic framework of MacCartney and Manning (2009) and find that less than 10% of entries represent an exact equivalence relation. Another task that moves away from the strict definition of semantic equivalence and seeks to understand the degree of similarity between two texts is that of *semantic textual similarity* (STS). STS assigns a real-valued score to a sentence pair that captures a notion of graded similarity (ranging from exact equivalence to complete dissimilarity). (Agirre et al., 2012, 2013; Corley and Mihalcea, 2005). Although STS provides a more fine-grained notion of similarity than paraphrase detection, it does not give insights into the nature of the textual difference. One framework that addresses this problem is *textual entailment* (TE) (Dagan et al., 2005). Recognizing TE is the task of determining whether the meaning of one text can be inferred (entailed) from another given text. Unlike the two framing we discussed above, TE is intrinsically asymmetric, e.g., the text *"A mean leans over a pickup track"* entails the text *"A man is touching a truck"*, but the reverse is not true.

Detecting Semantic Relations Models based on neural networks have been proposed for both monolingual and cross-lingual versions of the tasks (He et al., 2015; Rocktäschel et al., 2016; Tai et al., 2015). Recently, advances in language modeling based on deep transformer models led to adopting a single model architecture for modeling all the tasks above. To that end, a large transformer model is first pre-trained on large amounts of monolingual data using a masked language modeling objective. These pre-trained sentence representations then lead to improved performance on downstream tasks when fine-tuned on even small

amounts of task-specific training data (Devlin et al., 2019). Crucially, multilingual language models (Conneau and Lample, 2019) are pre-trained on huge amounts of unlabeled data from multiple languages with the hope that low-resource languages can benefit from high-resource languages. This enables their direct application in few- or zero-shot settings where labeled training data is scarce to non-existent.

MONO-LINGUAL	
Semantic Textual Similarity	English (Agirre et al., 2012, 2013, 2014, 2015, 2016), Spanish (Agirre et al., 2015), Arabic (Cer et al., 2017)
Paraphrase Detection	English, French, Spanish, German, Chinese, Japanese, Korean (Yang et al., 2019)
Textual Inference	English, French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili, Urdu (Conneau et al., 2018)
CROSS-LINGUAL	
Semantic Textual Similarity	Spanish-English (Agirre et al., 2016; Cer et al., 2017), Arabic-English (Cer et al., 2017)
Textual Inference	Spanish-English, German-English, Italian-English, French-English (Negri et al., 2012)
Semantic Divergences	French-English (Vyas et al., 2018a; Zhai et al., 2018), Spanish-English (Wein and Schneider, 2021), Chinese-English (Zhai et al., 2018)

Table 2.4: Language coverage in cross-sentence semantic relation tasks.

Chapter 3: Detecting Fine-Grained Cross-Lingual Semantic Divergences in the Wild[1]

In our first piece of work, we start our exploration by asking: ***How frequent are fine-grained semantic divergences in parallel texts?*** Our goal in this chapter is to address challenges in detection of fine meaning mismatches within bitexts in two settings: *human annotation* and *automatic prediction* of cross-lingual semantic divergences. More importantly, we address semantic divergences in the wild, i.e., without restricting ourselves to studying specific divergent phenomena or types. The latter poses a challenge as meaning mismatches in bitexts might vary across two dimensions. First, they might exhibit differences in their *nature* e.g., a word/phrase within a text corresponds to a non-equivalent translation in the counterpart text (see colored spans in example (A) below) *vs.* a phrase in one text corresponds to a piece of information missing from its counterpart (see colored span in example (B) below). Second, they might exhibit differences in their *granularity* e.g.; divergence might be attributed to a few words (see colored spans in examples (A) & (B) below) *vs.* the whole text (see unrelated texts in example (C) below).

[1] **Eleftheria Briakou** & Marine Carpuat. "Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank". In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

(A) EN The stores also offer class sessions as well as individual appointments.
FR Les magasins offrent également des séances en classe ainsi que des formations.
GLOSS *The stores also offer classroom sessions as well as training.*

(B) EN Tarek El Taib is a Libyan football midfielder.
FR Tarik El Taib né à Tripoli le 28 février 1977, est un footballeur libyen qui joue au poste de milieu de terrain.
GLOSS *Tarik El Taib born in Tripoli on February 28, 1977, is a Libyan footballer who plays as a midfielder.*

(C) EN In this capacity, she also attends the Health and Wellbeing Board, Norfolk.
FR Il a également eu des missions de conseil pour le Agence norvégienne de la santé et des services sociaux.
GLOSS *He has also had consultancy assignments for the Norwegian Agency for Health and Social Services.*

As a result, annotation of fine-grained semantic divergences is a hard task that requires protocols that establish common ground definitions across different annotators, especially when targeting non-experts (i.e., annotators with no linguistic background). To address this problem, we propose to move beyond a classification-based framework (i.e., directly annotate bitexts as *divergent vs. equivalent*)—used for annotation of coarse-grained semantic divergences (Vyas et al., 2018a), to a novel divergence annotation protocol that exploits rationales (Zaidan et al., 2007b)—highlights of substrings within the bitext that indicate divergences, to improve annotator agreement. Using our proposed protocol, we annotate 1,000 English-French parallel texts and find those fine-grained divergences not only exist in the wild but are also frequent, comprising 40% of our annotations. In Section §3.1, we describe our protocol along with our collected dataset, **R**ationalized **E**nglish-**Fre**nch **S**emantic **D**ivergences corpus (REFRESD), a dataset annotated with fine-grained semantic divergences at a sentence- and token-level.

After providing evidence that fine-grained divergences exist and can be detected by human annotations via carefully designing annotation protocols, we move to detect them via computational approaches. Given that annotation of divergences is expensive and hard to scale, the main challenge we face in automatically detecting them is the lack of supervised data for the task at hand. In Section §3.2, we address this limitation via introducing Divergent mBERT, a BERT-based model that detects fine-grained semantic divergences without supervision by learning to rank synthetic divergences of varying granularity. Concretely,

we automatically generate synthetic divergences by explicitly considering diverse types of fine-grained semantic divergences in bilingual text. Experiments on REFRESD show that our model distinguishes semantically equivalent from divergent examples much better than a strong sentence similarity baseline and that unsupervised token-level divergence tagging offers promise to refine distinctions among divergent instances.

3.1 Annotating Fine-Grained Cross-Lingual Semantic Divergences with Rationales

We design an annotation protocol to encourage annotators' sensitivity to semantic divergences other than misalignments without requiring expert knowledge beyond competence in the languages of interest. We use two strategies for this purpose: (1) we explicitly introduce distinct divergence categories for unrelated sentences and sentences that overlap in meaning, and (2) we ask for annotation rationales (Zaidan et al., 2007b) by requiring annotators to highlight tokens indicative of meaning differences in each sentence-pair. Thus, our approach strikes a balance between coarsely annotating sentences with binary distinctions that are fully based on annotators' intuitions (Vyas et al., 2018a) and exhaustively annotating all spans of a sentence-pair with fine-grained labels of translation processes (Zhai et al., 2018). We apply our annotation protocols to mined bitexts and contribute the **R**ationalized **E**nglish-**Fre**nch **S**emantic **D**ivergences (REFRESD) dataset, which consists of 1,039 English-French sentence-pairs annotated with sentence-level divergence judgments and token-level rationales. In the following subsections, we describe the annotation process and an analysis of the collected instances based on data statements protocols described in Bender and Friedman (2018); Gebru et al. (2018).[2] Concretely, we describe the annotation protocol (i.e., *what are the*

[2]The full description of our annotated dataset based on Data Statements (Bender and Friedman, 2018) can be found here: https://elbria.github.io/post/refresd/files/REFreSD_Data_Statements.pdf and the DataSheet (Gebru et al., 2018) for our dataset can be found here: https:

instructions of the annotation task?) in §3.1.1, the annotator demographics (i.e., *who was involved in the collection process?*) in §3.1.2, the curation rationale (i.e., *which texts were included and what were the goals in selecting texts?*) in §3.1.3, the quality control methods (i.e., *what procedures were followed to ensure good quality of annotations?*) in §3.1.4, the inter-annotator agreement (i.e., *how well multiple annotators can make the same annotation decision?*) in §3.1.5, and finally, we present annotation results in §3.1.6.

3.1.1 Annotation Protocol

An annotation instance consists of an English-French sentence pair. Bilingual participants are asked to read them both and highlight tokens in each sentence that convey meaning not found in the other language. For each highlighted span, they pick whether this span conveys added information ("**ADDED**"), information that is present in the other language but not an exact match ("**CHANGED**"), or some other type ("**OTHER**"). Those fine-grained classes are added to improve consistency across annotators and encourage them to read and compare the text closely. Finally, participants are asked to make a sentence-level judgment by selecting one of the following classes: "**NO MEANING DIFFERENCE**", "**SOME MEANING DIFFERENCE**", "**UNRELATED**". Participants are not given specific instructions on how to use span annotations to make sentence-level decisions. Furthermore, participants have the option of using a text box to provide any comments or feedback on the example and their decisions. A screenshot of the annotation interface is presented in Figure 3.1 while the exact guidelines presented to participants are included below:

> You are asked to compare the meaning of English and French text excerpts. You will be presented with one pair of texts at a time (about a sentence in English and a sentence in

`//elbria.github.io/post/refresd/files/REFreSD_Datasheet.pdf`.

French). For each pair, you are asked to do the following:

1 Read the two sentences carefully. Since the sentences are provided out of context, your understanding of content should only rely on the information available in the sentences. There is no need to guess what additional information might be available in the documents the excerpts come from.

2 Highlight the text spans that convey different meanings in the two sentences. After highlighting a span of text, you will be asked to further characterize it as:

> **ADDED** *the highlighted span corresponds to a piece of information that* ***does not exist*** *in the other sentence*
> **CHANGED** *the highlighted span corresponds to a piece of information that exists in the other sentence, but* ***their meaning is not the exact same***
> **OTHER** *none of the above holds*

You can highlight as many spans as needed. You can **optionally** provide an explanation for your assessment in the text form under the Notes section (e.g., literal translation of idiom)

3 Compare the meaning of the two sentences by picking one of the three classes:

UNRELATED *The two sentences are completely unrelated or have a few words in common but convey unrelated information about them*

SOME MEANING DIFFERENCE *The two sentences convey **mostly the same information, except differences for some details or nuances** (e.g., some information is added and/or missing on either or both sides; some English words have a more general or specific translation in French)*

NO MEANING DIFFERENCE *The two sentences have the **exact same meaning***

3.1.2 Annotator Demographics

We recruit 6 participants from the University of Maryland, College Park (UMD) educational institution. Recruitment for this project is driven by emails and flyers circulated via relevant UMD units, as well as through personal contacts in these units. Participants range in age from 20–25 years, including one man and five women. For each participant, we ensure they are proficient in both languages of interest: three of them self-report as English native speakers, one as a French native speaker, and two as bilingual English-French speakers. All non-bilingual L1-English speakers are graduate students in French Translation Studies, while the L1-French participant pursues a degree in Linguistics at UMD. Annotators are rewarded with Amazon gift cards at a rate of $2 per 10 examples, with bonuses of $5 and $10 for completing the first and additional sessions, respectively. We run 8 online annotation sessions. Each session consists of 120 instances, annotated by 3 participants, and lasts about 2 hours. Participants are allowed to take breaks during the process.

3.1.3 Curation rationale

Examples are drawn from the English-French section of the WikiMatrix corpus (Schwenk et al., 2021a). We chose this resource because (1) it is likely to contain diverse, interesting divergence types since it consists of mined parallel sentences of diverse topics which are not necessarily generated by (human) translations, and (2) Wikipedia and WikiMatrix are widely used resources to train semantic representations and perform cross-lingual transfer in NLP. We exclude obviously noisy samples by filtering out sentence pairs that a) are too short or too long, b) consist mostly of numbers, and c) have a small token-level edit difference. The filtered version of the corpus consists of $2{,}437{,}108$ sentence pairs.

3.1.4 Quality Control Methods

We implement quality control strategies at every step. We build a dedicated task interface using the BRAT annotation toolkit (Stenetorp et al., 2012) (Figure 3.1). We recruit participants from an educational institution and ensure they are proficient in both languages of interest. Specifically, participants are either bilingual speakers or graduate students pursuing a Translation Studies degree. We run a pilot study where participants annotate a sample containing both duplicated and reference sentence pairs previously annotated by one of the authors. All annotators are found to be internally consistent on duplicated instances and agree with the reference annotations more than 60% of the time. We solicit feedback from participants to finalize the instructions.

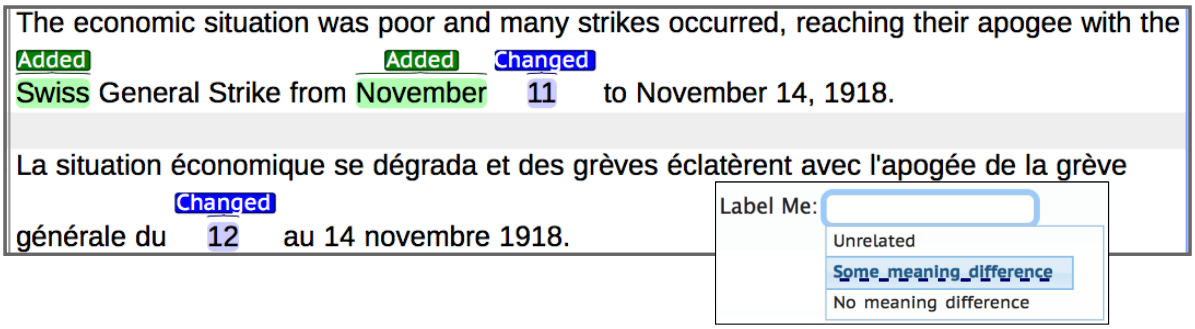


Figure 3.1: Screenshot of an example annotated instance.

3.1.5 Inter-annotator Agreement

We compute IAA for sentence-level annotations, as well as for the token and span-level rationales (Table 3.1). We report 0.60 Krippendorf's α coefficient for sentence classes, which indicates a "moderate" agreement between annotators (Landis and Koch, 1977). This constitutes a significant improvement over the 0.41 and 0.49 reported agreement coefficients on crowdsourced annotations of equivalence vs. divergence English-French parallel sentences drawn from OpenSubtitles and CommonCrawl corpora by prior work (Vyas et al., 2018a).

Disagreements mainly occur between the "**NO MEANING DIFFERENCE**" and "**SOME MEANING DIFFERENCE**" classes, which we expect as different annotators might draw the line between which differences matter differently. We only observed 3 examples where all 3 annotators disagreed (tridisagreements), which indicates that the "**UNRELATED**" and "**NO MEANING DIFFERENCE**" categories are more clear-cut. The rare instances with tridisagreements and bidisagreements—where the disagreement spans the two extreme classes—were excluded from the final dataset.

Quantifying agreement between rationales requires different metrics. At the span level; we compute macro F1 score for each sentence-pair following DeYoung et al. (2020a), where we treat one set of annotations as the reference standard and the other set as predictions.

Granularity	Method	INTER-ANNOTATOR AGREEMENT
Sentence	Krippendorf's α	0.60
Span	macro F1	45.56 ± 7.60
Token	macro F1	33.94 ± 8.24

Table 3.1: Inter-annotator agreement measured at different levels of granularities (i.e., sentence, span, and token) for the REFRESD dataset.

We count a prediction as a match if its token-level Intersection Over Union (IOU) with any of the reference spans overlaps by more than some threshold (here, 0.5). We report average span-level and token-level macro F1 scores computed across all different pairs of annotators. Average statistics indicate that our annotation protocol enabled the collection of a high-quality dataset.

3.1.6 Annotation Results: The REFRESD dataset

We aggregate sentence-level annotations by majority vote, yielding 252, 418, and 369 instances for the "**UNRELATED**", "**SOME MEANING DIFFERENCE**", and "**NO MEANING DIFFERENCE**" classes, respectively. In other words, 64% of samples are divergent, and 40% of samples contain fine-grained meaning divergences, confirming that divergences vary in granularity and are too frequent to be ignored even in a corpus viewed as parallel.

Table 3.2 includes examples of annotated instances drawn from REFRESD, corresponding to high sentence-level annotator agreement. In principle, we observe that span-level disagreements could occur within instances that are assigned the same sentence annotation, e.g., instances (E) & (F). This reveals that the reasoning process behind detecting divergences at a sentence level might differ across annotators, and some types of divergences might be more difficult to detect than others. Manual inspection of instances that exhibit high agreement at both a sentence and a span level suggests that some examples constitute clear-cut divergences. For instance, in annotating the instance (B) as containing **SOME**

MEANING DIFFERENCE, all annotators provide the same highlighted span, i.e., *"en réponse à une baisse notable des ventes"*, as containing **ADDED** information. On the other hand, annotators provide different highlights for the instance (E) that is unanimously identified as belonging to the **SOME MEANING DIFFERENCE** class. For instance, the French segment *"renouvelé la série"* (i.e., renewed the season), which is mapped to the more general meaning of *"ordered a season"* in the English text, is highlighted by one of the three annotators. Those examples reveal that span-level differences between annotators are not necessarily random attention slips but rather could result from different intuitions on where to draw the line between divergence and equivalence definitions in cases of subtle meaning differences. As a result, our REFRESD dataset could be used to shed light on the difficulty and the nature of the divergences detection task by studying the raw disagreements closely.

Furthermore, annotators' disagreements are also observed at the sentence level. Table 3.3 presents examples from REFRESD that correspond to bi-disagreements across closely related classes. Again, we observe that disagreements reflect the difficulty of the task due to the wide spectrum of different divergences types and granularities. For instance, in annotating the (G) instance, the **ADDED** information "further refined" in the English text, is only considered as a difference that contributes to semantic divergences by 1 out of the 3 annotators, while the rest of them consider it irrelevant to the meaning of sentences, i.e., annotating the instance as containing **NO MEANING DIFFERENCE**. Similarly, disagreements arise between the **SOME MEANING DIFFERENCE** and **UNRELATED** classes. Those disagreements are primarily observed in cases where there are coarse differences in content, yet sentences might still share some content in common at the subsentential level, e.g., see (I).

(A)	CLASS	NO MEANING DIFFERENCE
	EN	The plan was revised in 1916 to concentrate the main US naval fleet in New England, and from there defend the US from the German navy.
	FR	Le plan fut révisé en 1916 pour concentrer le gros de la flotte navale américaine en Nouvelle-Angleterre, et à partir de là, défendre les États-Unis contre la marine allemande.
	GLOSS	The plan was revised in 1916 to concentrate the bulk of the American naval fleet in New England, and from there defend the United States against the German Navy.
(B)	CLASS	SOME MEANING DIFFERENCE
	EN	After an intermediate period during which Stefano Piani edited the stories, in 2004 a major rework of the series went through.
	FR	Après une période intermédiaire pendant laquelle Stefano Piani édita les histoires, une refonte majeure de la série fut faite en 2004 en réponse à une baisse notable des ventes.
	GLOSS	After an interim period during which Stefano Piani edited the stories, a major overhaul of the series was made in 2004 in response to a noticeable drop in sales.
(C)	CLASS	UNRELATED
	EN	To reduce vibration, all helicopters have rotor adjustments for height and weight.
	FR	En vol, le régime du compresseur Tous les compresseurs ont un taux de compression lié à la vitesse de rotation et au nombre d'étages.
	GLOSS	In flight, the compressor speed All compressors have a compression ratio linked to the speed of rotation and the number of stages.
(D)	CLASS	NO MEANING DIFFERENCE
	EN	One can see two sunflowers on the main façade and three smaller ones on the first floor above ground just above the entrance arcade.
	FR	On remarquera deux tournesols sur la façade principale et trois plus petits au premier étage au-dessus des arcades d'entrée.
	GLOSS	One will notice two sunflowers on the main facade and three smaller ones on the first floor above the entrance arcades.
(E)	CLASS	SOME MEANING DIFFERENCE
	EN	On November 10, 2014, CTV ordered a fourth season of Saving Hope that consisted of eighteen episodes, and premiered on September 24.
	FR	Le 10 novembre 2014, CTV a renouvelé la série pour une quatrième saison de 18 épisodes diffusée depuis le 24 septembre 2015.
	GLOSS	On November 10, 2014, CTV renewed the series for an 18-episode fourth season that aired since september 24, 2015.
(F)	CLASS	UNRELATED
	EN	He talks about Jay Gatsby, the most hopeful man he had ever met.
	FR	Il côtoie notamment Giuseppe Meazza qui dira de lui Il fut le joueur le plus fantastique que j'aie eu l'occasion de voir.
	GLOSS	He is especially close to Giuseppe Meazza who will say of him He was the most fantastic player I have ever had the chance to see.

Table 3.2: REFRESD examples corresponding high sentence-level inter-annotator agreement, i.e., all 3 annotators agree.

(G)	CLASS	NO MEANING DIFFERENCE (n=2); SOME MEANING DIFFERENCE (n=1)
	EN	Nine of these revised BB LMs were built by Ferrari in 1979, while a further refined series of sixteen were built from 1980 to 1982.
	FR	Neuf de ces BB LM révisées furent construites par Ferrari en 1979, tandis qu'une série de seize autres furent construite entre 1980 et 1982.
	GLOSS	Nine of these revised BB LMs were built by Ferrari in 1979, while a series of sixteen others were built between 1980 and 1982.
(H)	CLASS	SOME MEANING DIFFERENCE (n=2); NO MEANING DIFFERENCE (n=1)
	EN	From 1479, the Counts of Foix became Kings of Navarre and the last of them, made Henri IV of France, annexed his Pyrrenean lands to France.
	FR	À partir de 1479, le comte de Foix devient roi de Navarre et le dernier d'entre eux, devenu Henri IV, roi de France en 1607, annexe ses terres pyrénéennes à la France.
	GLOSS	From 1479, the Count of Foix became King of Navarre and the last of them, who became Henri IV, King of France in 1607, annexed his Pyrenean lands to France.
(I)	CLASS	UNRELATED (n=2) SOME MEANING DIFFERENCE (n=1);
	EN	The operating principle was the same as that used in the Model 07/12 Schwarzlose machine gun used by Austria-Hungary during the First World War.
	FR	Le Skoda 100 mm modèle 1916 était un obusier de montagne utilisé par l'Autriche-Hongrie pendant la Première Guerre mondiale.
	GLOSS	The Skoda 100mm Model 1916 was a mountain howitzer used by Austria-Hungary during World War I.

Table 3.3: REFRESD examples corresponding to medium sentence-level inter-annotator agreement i.e., 2 out of 3 annotators voted for the sentence-level majority class. Disagreements span closely related classes, while n gives the number of annotators voted for each class.

3.1.7 Summary

We presented an annotation protocol for collecting semantic divergence annotations that is based on a two-step approach that uses span-level rationales to guide the final clustering of bitexts in three classes, namely **NO MEANING DIFFERENCE**, **SOME MEANING DIFFERENCE**, and **UNRELATED**. We applied our protocol on WikiMatrix English-French bitexts via carefully recruiting bilingual speakers and collected annotations on $1{,}047$ bitexts that we released with our REFRESD dataset. Overall, our annotation exercise shows that fine-grained divergences (1) can be reliably annotated under several steps of quality control and (2) are frequent even in a dataset that is perceived as parallel. At the same time, we show that manual annotation of fine-grained semantic divergences is hard to scale. Concretely, we found that collecting divergence annotations in REFRESD is time-consuming, i.e., required ~ 54 hours of annotations, and expensive, i.e., $\$780$ of annotators' compensation. In our next subsection, we ask: ***How can we detect fine-grained semantic divergences at scale?*** In answering this question, we take a computational approach that draws from our annotation findings but, crucially, does not require access to human-labeled data for training or development.

3.2 Unsupervised Detection of Fine-Grained Cross-Lingual Semantic Divergences

In this subsection, we propose an unsupervised divergence detection model that addresses a binary classification task. Concretely, given two inputs: an English sentence $\mathbf{x_e}$ and a French sentence $\mathbf{x_f}$, our model outputs a binary prediction, i.e., equivalence vs. divergence. Unlike the prior work of Vyas et al. (2018a), who use the same task definition to detect coarse-grained semantic divergences in bitexts, we aim instead at capturing both fine- and

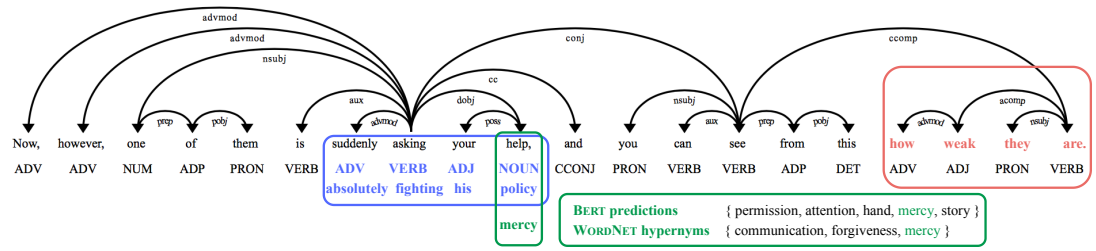
coarse-grained divergence under the divergence class. In doing so, we explicitly consider diverse types of semantic divergences in bilingual texts when modeling and evaluating divergence detection. In the following subsections, we present our model in §3.2.1, the experimental conditions in §3.2.2, and the experimental results in §3.2.3.

3.2.1 Divergent mBERT: Learning to Rank Synthetic Contrastive Divergences

We introduce a model based on multilingual BERT (mBERT) to distinguish divergent from equivalent sentence-pairs (¶**Fine-tuning mBERT**). In the absence of annotated training data, we derive synthetic divergent samples from parallel corpora (¶**Generating Synthetic Divergences**) and train via learning to rank to exploit the diversity and varying granularity of the resulting samples (¶**Learning to Rank Contrastive Samples**).

Fine-tuning mBERT Given the success of multilingual masked language models like mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), and XLM-R (Conneau et al., 2020) on cross-lingual understanding tasks, we build our classifier on top of multilingual BERT in a standard fashion: we create a sequence $\mathbf{x}$ by concatenating $\mathbf{x_e}$ and $\mathbf{x_f}$ with helper delimiter tokens: $\mathbf{x} = (\text{[CLS]}, \mathbf{x_e}, \text{[SEP]}, \mathbf{x_f}, \text{[SEP]})$. The [CLS] token encoding serves as the representation for the sentence-pair $\mathbf{x}$, passed through a feed-forward layer network F to get the score $F(\mathbf{x})$. Finally, we convert the score $F(\mathbf{x})$ into the probability of $\mathbf{x}$ belonging to the equivalent class.

Generating Synthetic Divergences Drawing from our annotation findings in §3.1, we devise three ways of creating training instances that mimic fine-grained divergences of *different nature* and *varying granularity* by perturbing seed equivalent samples from parallel corpora (Table 3.4):



Seed Equivalent Sample

Now, however, one of them is suddenly asking your help, and you can see from this how weak they are.
Maintenant, cependant, l'un d'eux vient soudainement demander votre aide et vous pouvez voir à quel point ils sont faibles.

Subtree Deletion

Now, however, one of them is suddenly asking your help, and you can see from this.
Maintenant, cependant, l'un d'eux vient soudainement demander votre aide et vous pouvez voir à quel point ils sont faibles .

Phrase Replacement

Now, however, one of them is absolutely fighting his policy , and you can see from this how weak they are.
Maintenant, cependant, l'un d'eux vient soudainement demander votre aide et vous pouvez voir à quel point ils sont faibles.

Lexical Substitution

Now, however, one of them is suddenly asking your mercy , and you can see from this.
Maintenant, cependant, l'un d'eux vient soudainement demander votre aide et vous pouvez voir à quel point ils sont faibles.

Table 3.4: Starting from a seed equivalent parallel sentence-pair, we create three types of divergent samples of varying granularity by introducing the highlighted edits.

1. **Subtree Deletion** We mimic semantic divergences due to content included only in one language by deleting a randomly selected subtree in the dependency parse of the English sentence or French words aligned to English words in that subtree. We use subtrees that are not leaves and that cover less than half of the sentence length. Durán et al. (2014); Cardon and Grabar (2020) successfully use this approach to compare sentences in the same language.
2. **Phrase Replacement** Following Pham et al. (2018), we introduce divergences that mimic phrasal edits or mistranslations by substituting random source or target sequences by another sequence of words with matching POS tags (to keep generated sentences as grammatical as possible).
3. **Lexical Substitution** We mimic *particularization* and *generalization* translation operations (Zhai et al., 2019b) by substituting English words with hypernyms or hyponyms from WordNet. The replacement word is the highest scoring WordNet

candidate in context, according to a BERT language model (Zhou et al., 2019b; Qiang et al., 2019).

We call all these divergent examples **contrastive** because each divergent example contrasts with a specific equivalent sample from the seed set. The three sets of transformation rules above create divergences of varying granularity and create an **implicit ranking over divergent examples** based on the range of edit operations, starting from a single token with lexical substitution to local short phrases for phrase replacement and up to half the words in a sentence when deleting subtrees.

Learning to Rank Contrastive Samples We train the Divergent mBERT model by learning to rank synthetic divergences. Instead of treating equivalent and divergent samples independently, we exploit their contrastive nature by explicitly pairing divergent samples with their seed equivalent sample when computing the loss. Intuitively, *lexical substitution* samples should rank higher than *phrase replacement* and *subtree deletion* and lower than seed *equivalents*: we exploit this intuition by enforcing a margin between the scores of increasingly divergent samples.

Formally, let $\mathbf{x}$ denote an English-French sentence-pair and $\mathbf{y}$ a contrastive pair, with $\mathbf{x} > \mathbf{y}$ indicating that the divergence in $\mathbf{x}$ is finer-grained than in $\mathbf{y}$. For instance, we assume that $\mathbf{x} > \mathbf{y}$ if $\mathbf{x}$ is generated by lexical substitution and $\mathbf{y}$ by subtree deletion.

At training time, given a set of contrastive pairs $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})\}$, the model is trained to rank the score of the first instance higher than the latter by minimizing the following margin-based loss

$$\mathcal{L}_{\text{sent}} = \frac{1}{|\mathcal{D}|}\Big(\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \max\{0, \xi - F(\mathbf{x}) + F(\mathbf{y})\}\Big) \tag{3.1}$$

where ξ is a hyperparameter margin that controls the score difference between the sentence pairs $\mathbf{x}$ and $\mathbf{y}$. This ranking loss has proved useful in supervised English semantic analysis

tasks (Li et al., 2019), and we show that it also helps with our cross-lingual synthetic data.

3.2.2 Experimental Conditions

Data We normalize English and French text in WikiMatrix consistently using the Moses toolkit (Koehn et al., 2007) and tokenize into subword units using the "BertTokenizer". Specifically, our pre-processing pipeline consists of a) replacement of Unicode punctuation, b) normalization of punctuation, c) removal of non-printing characters, and d) tokenization.[3] We align English to French bitext using the Berkeley word aligner.[4] We filter out obviously noisy parallel sentences, as described in Section 3.1.3, Curation Rationale. The top 5,500 samples ranked by LASER similarity score are treated as (noisy) equivalent samples and seed the generation of synthetic divergent examples.[5] We split the seed set into 5,000 training instances and 500 development instances consistently across experiments.

Implementation Our models are based on the HuggingFace transformer library (Wolf et al., 2019).[6] We fine-tune the "BERT-Base Multilingual Cased" model (Devlin et al., 2019),[7] and perform a grid search on the margin hyperparameter, using the synthetic development set.

Baselines We compare our Divergent mBERT model with prior work (LASER baseline) and with several ablation variants of our approach that establish the importance of our design decisions (Divergent mBERT variants):

- LASER baseline This baseline distinguishes equivalent from divergent samples via a threshold on the LASER score. We use the same threshold as Schwenk et al. (2021a),

[3] https://github.com/facebookresearch/XLM/blob/master/tools/tokenize.sh
[4] https://code.google.com/archive/p/berkeleyaligner
[5] https://github.com/facebookresearch/LASER/tree/master/tasks/WikiMatrix
[6] https://github.com/huggingface/transformers
[7] https://github.com/google-research/bert

who show that training Neural Machine Translation systems on WikiMatrix samples with LASER scores higher than 1.04 improves BLEU. Preliminary experiments suggest that tuning the LASER threshold does not improve classification. to distinguish coarse-grained divergence vs. equivalence.

- Divergent mBERT variants We compare Divergent mBERT trained by learning to rank contrastive samples (Section 3.2.1) with ablation variants.

  To test the impact of contrastive training samples, we fine-tune Divergent mBERT using

  1. the Cross-Entropy (**CE**) loss on randomly selected synthetic divergences;
  2. the **CE** loss on paired equivalent and divergent samples, treated as independent;
  3. the proposed training strategy with a **Margin** loss to explicitly compare contrastive pairs.

  Given the fixed set of seed equivalent samples (Section 3.2.2, Data), we vary the combinations of divergent samples:

  1. **Single divergence type** we pair each seed equivalent with its corresponding divergent of that type, yielding a single contrastive pair;
  2. **Balanced sampling** we randomly pair each seed equivalent with one of its corresponding divergent types, yielding a single contrastive pair;
  3. **Concatenation** we pair each seed equivalent with one of each synthetic divergence type, yielding four contrastive pairs;
  4. **Divergence ranking** we learn to rank pairs of close divergence types: equivalent vs. lexical substitution, lexical substitution vs. phrase replacement, or subtree deletion yielding four contrastive pairs. For lexical substitution, we mimic both generalization **and** particularization.

Evaluation We evaluate all models on our new REFRESD dataset using Precision, Recall, F1 for each class, and Weighted overall F1 score as computed by *scikit-learn* (Pedregosa et al., 2011).[8]

3.2.3 Findings

Table 3.5 presents experimental results on our REFRESD dataset. All Divergent mBERT models outperform the LASER baseline by a large margin. The proposed training strategy performs best, improving over LASER by 31 F1 points. Ablation experiments and analysis further show the benefits of diverse contrastive samples and learning to rank.

Contrastive Samples With the CE loss, independent contrastive samples improve over randomly sampled synthetic instances overall ($+8.7$ F1+ points on average) at the cost of a smaller drop for the divergent class (-5.3 F1- points) for models trained on a single type of divergence. Using the margin loss helps models recover from this drop.

Divergence Types All types improve over the LASER baseline. When using a single divergence type, *Subtree Deletion* performs best, even matching the overall F1 score of a system trained on all types of divergences (*Balanced Sampling*). Training on the *Concatenation* of all divergence types yields poor performance. We suspect that the model is overwhelmed by negative instances at training time, which biases it toward predicting the divergent class too often and hurting F1+ score for the equivalent class.

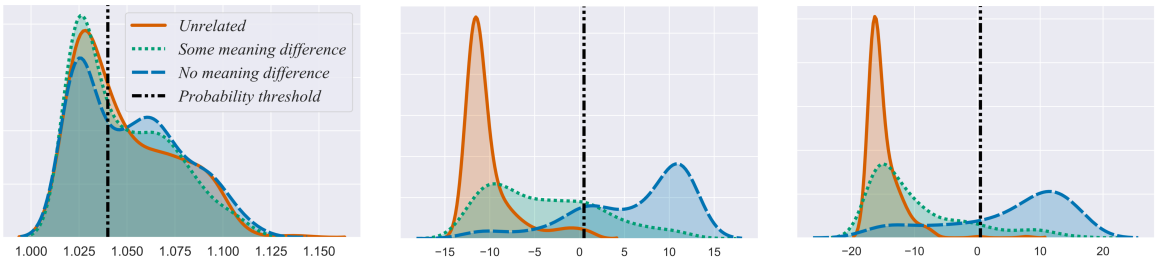
Divergence Ranking How does divergence ranking improve predictions? Figure 3.2 shows model score distributions for the 3 classes annotated in REFRESD. *Divergence Ranking* particularly improves divergence predictions for the "Some meaning difference"

[8] https://scikit-learn.org

class: the score distribution for this class is more skewed toward negative values than when training on contrastive *Subtree Deletion* samples.

			Equivalents			*Divergents*			*All*		
Synthetic	**Loss**	**Contrastive**	**P+**	**R+**	**F1+**	**P-**	**R-**	**F1-**	**P**	**R**	**F1**
Phrase Replacement	CE	✗	70	56	62	78	87	82	75	76	75
		✓	61	81	69	87	71	78	78	75	75
	Margin	✓	70	76	73	86	82	84	80	80	80
Subtree Deletion	CE	✗	81	50	62	77	93	85	78	78	77
		✓	64	84	72	89	74	81	80	77	78
	Margin	✓	70	83	76	90	81	85	83	82	82
Lexical Substitution	CE	✗	65	53	57	76	84	80	72	73	72
		✓	55	81	66	86	64	73	75	70	71
	Margin	✓	57	75	65	83	70	76	74	72	72
Balanced	CE	✗	76	42	54	74	93	83	75	75	73
		✓	73	73	73	85	85	85	81	81	81
	Margin	✓	76	73	75	85	87	86	82	82	82
Concatenation	CE	✗	62	32	42	70	89	79	67	69	66
		✓	73	55	63	78	89	83	76	77	76
	Margin	✓	84	59	70	81	94	87	82	82	81
Divergence Ranking	Margin	✓	82	72	**77**	86	91	**88**	84	85	**84**
LASER baseline			38	58	46	68	48	57	57	52	53

Table 3.5: Intrinsic evaluation of Divergent mBERT and its ablation variants on the REFRESD dataset. We report Precision (P), Recall (R), and F1 for the equivalent (+) and divergent (-) classes separately, as well as for both classes (*All*). *Divergence Ranking* yields the best F1 scores across the board.



(a) LASER (b) *Subtree Deletion* (c) *Divergence Ranking*

Figure 3.2: Score distributions assigned by different models to sentence-pairs of REFRESD. *Divergence Ranking* scores for the "Some meaning difference" class are correctly skewed more toward negative values.

3.2.4 Summary

We show that explicitly considering diverse semantic divergence types benefits not only the *annotation* (§3.1) but also the *prediction* of divergences between texts in different languages. We show that these divergences can be detected by a mBERT model fine-tuned without annotated samples by learning to rank synthetic divergences of varying granularity. While we cast divergence detection as binary sentence-level classification in this subsection, human judges separated divergent samples into "Unrelated" and "Some meaning difference" classes in the REFRESD dataset while providing token-level divergence assessments in the form of rationales. In the next subsection, we ask: ***Can we make finer-grained divergence predictions beyond a binary classification framework?***

3.3 Unsupervised Finer-Grained Divergence Detection via Token Tagging

In this subsection, we target finer-grained divergence detection under two different task formulations: first, we explore whether divergences can be automatically detected at a token level, and then we test the hypothesis that token-level divergence predictions can help discriminate between divergence granularities at the sentence level, inspired by humans' use of rationales to ground sentence-level judgments. In the following subsections, we present an extension of Divergent mBERT for token-tagging in §3.3.1, the experimental conditions in §3.3.2, and the experimental results on token labeling and 3-way sentence classification in §3.3.3.

3.3.1 Divergent mBERT for Token Tagging

We introduce an extension of Divergent mBERT, which, given a bilingual sentence pair, produces a) a sentence-level prediction of equivalence vs. divergence and b) a sequence of

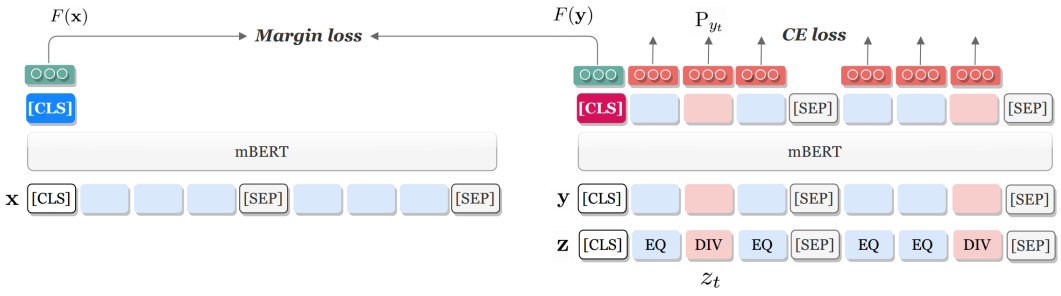


Figure 3.3: Divergent mBERT training strategy: given a triplet (x,y,z), the model minimizes the sum of a margin-based loss via ranking a contrastive pair $\mathbf{x} > \mathbf{y}$ and a token-level cross-entropy loss on sequence labels $\mathbf{z}$.

EQ/DIV labels for each input token. EQ and DIV refer to token-level tags of equivalence and divergence, respectively.

Motivated by annotation rationales, we adopt a multi-task framework to train our model on a set of triplets $\mathcal{D}' = \{(\mathbf{x}, \mathbf{y}, \mathbf{z})\}$, still using only synthetic supervision (Figure 3.3). As in Section 3.2.1, we assume $\mathbf{x} > \mathbf{y}$, while $\mathbf{z}$ is the sequence of labels for the second encoded sentence pair $\mathbf{y}$, such that, at time t, $z_t \in$ {EQ,DIV} is the label of y_t. Since Divergent mBERT operates on sequences of subwords, we assign an EQ or DIV label to a word token if at least one of its subword units is assigned that label.

For the token prediction task, the final hidden state h_t of each y_t token is passed through a feed-forward layer and a softmax layer to produce the probability P_{y_t} of the y_t token belonging to the EQ class. For the sentence task, the model learns to rank $\mathbf{x} > \mathbf{y}$, as in Section 3.2.1. We then minimize the sum of the sentence-level margin-loss and the average token-level cross-entropy loss ($\mathcal{L}_{\text{CE}}$) across all tokens of $\mathbf{y}$, as defined in Equation 3.2.

$$\mathcal{L} = \frac{1}{|\mathcal{D}'|}\Big(\sum_{(\mathbf{x},\mathbf{y},\mathbf{z})\in\mathcal{D}'} \big(\max\{0, \xi - F(\mathbf{x}) + F(\mathbf{y})\} + \frac{1}{|\mathbf{y}|}\sum_{t=1}^{|\mathbf{y}|} \mathcal{L}_{\text{CE}}(\mathrm{P}_{y_t}, z_t)\big)\Big) \tag{3.2}$$

Similar multi-task models have been used for Machine Translation Quality Estimation (Kim et al., 2019a,b), albeit with human-annotated training samples and a standard cross-entropy

loss for both word-level and sentence-level sub-tasks.

3.3.2 Experimental Conditions

Models We fine-tune the **multi-task** MBERT model that makes token and sentence predictions jointly, as described in Section 3.3. We contrast against a sequence labeling MBERT model trained independently with the CE loss (**Token-only**). Finally, we run a **random baseline** where each token is labeled EQ or DIV uniformly at random.

Training Data We tag tokens edited when generating synthetic divergences as DIV (e.g., highlighted tokens in Table 3.4), and others as EQ. Since edit operations are made on the English side, we tag-aligned French tokens using the Berkeley aligner.

Evaluation We expect token-level annotations in REFRESD to be noisy since they are produced as rationales for sentence-level rather than token-level tags. We, therefore, consider three methods to aggregate rationales into token labels: a token is labeled as DIV if it is highlighted by at least one (**Union**), two (**Pair-wise Union**), or all three annotators (**Intersection**). We report F1 on the DIV and EQ class and F1-Mul as their product for each of the three label aggregation methods.

3.3.3 Findings

Token Labeling We evaluate token labeling on REFRESD samples from the "Some meaning difference" class, where we expect the more subtle differences in meaning to be found and the token-level annotation to be most challenging (Table 3.6). Examples of Divergent MBERT's token-level predictions are given in the Appendix of Briakou and Carpuat (2020). The *Token-only* model outperforms the *Random Baseline* across all metrics, showing the benefits of training even with noisy token labels derived from rationales. *Multi-task* training

		Union			Pair-wise Union			Intersection		
Model	Multi-task	F1-DIV	F1-EQ	F1-Mul	F1-DIV	F1-EQ	F1-Mul	F1-DIV	F1-EQ	F1-Mul
Random Baseline		0.21	0.62	0.13	0.33	0.59	0.20	0.21	0.62	0.13
Token-only		0.39	0.77	0.30	0.46	**0.88**	0.41	0.46	**0.92**	0.42
Balanced	✓	0.41	0.77	0.32	0.46	0.87	0.40	0.43	0.91	0.40
Concatenation	✓	0.41	**0.78**	0.32	0.48	**0.88**	0.42	0.46	**0.92**	0.42
Divergence Ranking	✓	**0.45**	**0.78**	**0.35**	**0.51**	**0.88**	**0.45**	**0.49**	**0.92**	**0.45**

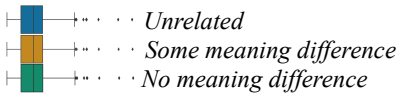
Table 3.6: Evaluation of different models on the token-level prediction task for the **"Some meaning difference"** class of REFRESD. *Divergence Ranking* yields the best results across the board.

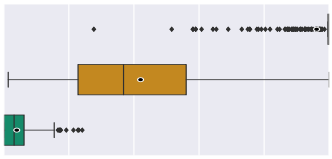
further improves over *Token-only* predictions for almost all different metrics. *Divergence Ranking* of contrastive instances yields the best results across the board. Also, on the auxiliary sentence-level task, the *Multi-task* model matches the F1 as the standalone *Divergence Ranking* model.

From Token to Sentence Predictions We compute the % of DIV predictions within a sentence-pair. The multi-task model makes more DIV predictions for the divergent classes as its % distribution is more skewed towards greater values (Figure 3.4 (d) vs. (e)). We then show that the % of DIV predictions of the *Divergence Ranking* model can be used as an indicator for distinguishing between divergences of different granularity: intuitively, a sentence pair with more DIV tokens should map to a coarse-grained divergence at a sentence-level. Table 3.7 shows that thresholding the % of DIV tokens could be an effective discrimination strategy, which we will explore further in future work.

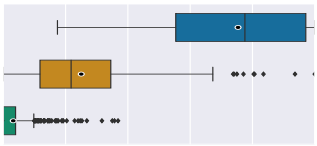
	UN			SD			
DIV	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**F1-all**
10%	48	**97**	64	66	51	57	59
20%	69	84	**76**	**83**	79	81	80
30%	82	63	71	81	**85**	**83**	**81**
40%	**94**	35	51	73	84	78	75

Table 3.7: "Some meaning difference" (SD) vs. "Unrelated" (UN) classification based on % of DIV labels.

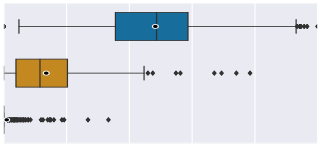




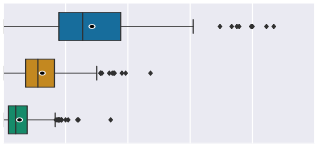
(a) *Union*



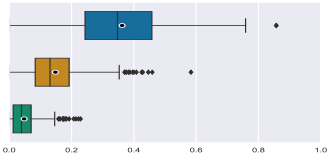
(b) *Pair-wise Union*



(c) *Intersection*



(d) *Token-only*



(e) *Divergence Ranking*

Figure 3.4: Percentage distributions of DIV tokens in REFRESD and DIV token predictions of two models: *Divergence Ranking* makes more DIV predictions compared to the *Token-only* model, enabling a better distinction between the divergent classes.

3.3.4 Summary

In this subsection, we draw inspiration from the rationale-based annotation process described in §3.1 and show that predicting token-level and sentence-level divergences jointly is a promising direction for further distinguishing between coarser and finer-grained divergences. After establishing that fine-grained semantic divergences exist and can be automatically detected without supervision, we now ask: ***How do fine-grained semantic divergences impact Neural Machine Translation?*** We aim to answer this question in the next chapter by contributing a controlled empirical analysis on several aspects of NMT models that are exposed to different types and amounts of fine-grained divergences at training time.

Chapter 4: Analyzing the Impact of Fine-grained Semantic Divergences on Neural Machine Translation[1]

While parallel texts are essential to Neural Machine Translation (NMT), the degree of parallelism varies widely across samples in practice, as discussed in Section §2.1.3. For instance (Figure 4.1), a French SOURCE could be paired with an exact translation into English (EQ), with a mostly equivalent translation where only a few tokens convey divergent meaning (fine-DIV), or with a semantically unrelated, noisy reference (coarse-DIV). Yet, as we discussed in Section §2.2.1, prior work treats parallel samples in a binary fashion: coarse-grained divergences are viewed as noise to be excluded from training (Koehn et al., 2018), whilst others are typically regarded as gold-standard equivalent translations. As a result, the impact of fine-grained divergences on NMT remains unclear.

[1] **Eleftheria Briakou** & Marine Carpuat. "Beyond Noise: Mitigating the Impact of Fine-grained Semantic Divergences on Neural Machine Translation." *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), 2021.*

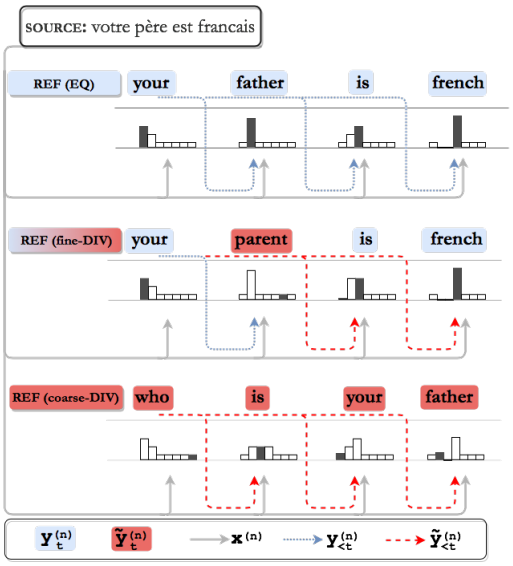


Figure 4.1: Equivalent vs. Divergent references on NMT training. Fine-grained divergences (i.e., REF (fine-DIV)) provide an imperfect yet potentially useful signal depending on the time step t.

Our second piece of work aims to understand the impact of fine-grained semantic divergences in NMT. In this direction, we first start by revisiting the NMT training assumptions (Section §4.1). Then, we contribute an analysis of how fine-grained divergences in training data affect NMT quality and confidence. Starting from a set of equivalent English-French WikiMatrix sentence pairs, we simulate divergences by gradually "corrupting" them with synthetic *fine-grained divergences* (Section §4.2). Following Khayrallah and Koehn (2018)—who, in contrast, study the impact of *noise* on MT—we control for different *types* of fine-grained semantic divergences and different *ratios* of equivalent vs. divergent data. Our findings (Section §4.4) indicate that these imperfect training references: hurt translation quality (as measured by BLEU and METEOR) once they overwhelm equivalents; output degenerated text more frequently; and increase the uncertainty of models' predictions.

4.1 Neural Machine Translation Training Assumptions

NMT models are typically trained to maximize the log-likelihood of the training data, $\mathcal{D} \equiv \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$, where $(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})$ is the n-th sentence pair consisting of sentences that are **assumed to be translations of each other**. Under this assumption, model parameters are updated to maximize the **token-level** cross-entropy loss:

$$\mathcal{J}(\theta) = \sum_{n=1}^{N} \sum_{t=1}^{T} \log p(y_t^{(n)} \mid \mathbf{y}_{<t}^{(n)}, \mathbf{x}^{(n)}; \theta) \tag{4.1}$$

In Figure 4.1, we illustrate how semantic divergences interact with NMT training. In the case of coarse divergences, both the prefixes $\widetilde{\mathbf{y}}_{t<1}^{(n)}$ and targets $\widetilde{y}_t^{(n)}$, yield a noisy training signal at each time step t, which motivates excluding them from the training pool entirely. In the case of fine-grained divergences, the assumption of *semantic equivalence* is only partially broken. Depending on the time step t, we might thus condition the prediction of the next token on partially corrupted prefixes, encourage the model to make a wrong prediction, or do a combination of the above. This suggests that fine-grained divergent samples provide a noisy yet potentially useful training signal depending on the time step. Meanwhile, fine-grained divergences increase uncertainty in the training data and, as a result, might impact models' confidence in their predictions, as noisy untranslated samples do (Ott et al., 2018).

4.2 Controlling the Types and Intensity of Divergences in Bitexts

We evaluate the impact of semantic divergences on NMT by injecting increasing amounts of synthetic divergent samples during training, following the methodology of Khayrallah and Koehn (2018) for noise. We focus on three types of divergences (i.e., PHRASE REPLACEMENT, LEXICAL SUBSTITUTION, and SUBTREE DELETION), which we found to be

frequent in parallel corpora, as discussed in Section §3. Synthetic divergent samples are automatically generated by corrupting semantically equivalent sentence pairs, following the methodology introduced in Section §3. Equivalents are identified by our Divergent mBERT classifier, and LEXICAL SUBSTITUTION, PHRASE REPLACEMENT, and SUBTREE DELETION divergent types are generated following the process described in Section §3.2.1. Having access to those 4 versions of the same corpus (one initial equivalent and three synthetic divergences), we mix equivalents and divergent pairs, introducing one type of divergence at a time. Tables 4.1 and 4.2 contain corpus statistics for the 3 versions of synthetic divergences we create, starting from EQUIVALENTS. LEXICAL SUBSTITUTION are sampled randomly from the pools of substitutions based on hypernyms and hyponyms.

WikiMatrix version	#Sents.	#Tokens	#Types	Length	%Corr
EQUIVALENTS	751,792	22,723,543	515,154	30.2	0%
SUBTREE DELETION	749,973	20,783,056	483,336	27.7	9.32%
PHRASE REPLACEMENT	750,527	22,735,143	475,567	30.3	16.11%
LEXICAL SUBSTITUTION (HYPERNYMS)	724,326	22,014,609	497,658	30.4	12.33%
LEXICAL SUBSTITUTION (HYPONYMS)	617,913	18,970,039	442,299	30.7	7.42%

Table 4.1: WikiMatrix statistics corresponding to extracted EQUIVALENTS and the fine-grained corruptions introduced in the synthetic setting (EN-side). %Corr denotes the average % of corrupted tokens in a sentence.

WikiMatrix Version	#Sents.	#Tokens	#Types	Length	%Corr
EQUIVALENTS	751,792	25,554,549	515,194	34.0	0%
SUBTREE DELETION	749,973	23,822,958	486,908	31.8	7.21%
PHRASE REPLACEMENT	750,527	25,554,549	515,194	34.0	12.74%
LEXICAL SUBSTITUTION (HYPERNYMS)	724,326	24,737,604	499,423	34.2	9.82%
LEXICAL SUBSTITUTION (HYPONYMS)	617,913	21,387,650	445,871	34.6	5.78%

Table 4.2: WikiMatrix statistics corresponding to extracted EQUIVALENTS and the fine-grained corruptions introduced in the synthetic setting (FR-side).

4.3 Experimental Conditions

Training Data We train our models on the parallel WikiMatrix French-English corpus (Schwenk et al., 2021a), which consists of sentence pairs mined from Wikipedia pages using language-agnostic sentence embeddings (LASER) (Artetxe and Schwenk, 2019). Previous annotations show that 40% of sentence pairs in a random sample contain fine-grained divergences (Briakou and Carpuat, 2020).

After cleaning noisy samples using simple rules (i.e., exclude pairs that are a) too short or too long, b) mostly numbers, c) almost copies based on edit distance), we extract *equivalent* samples using the Divergent mBERT model. Table 4.3 presents statistics on the extracted pairs, along with the corpus created if we threshold the LASER score at 1.04, as suggested by Schwenk et al. (2021a).

Corpus	**#Sentences**
WIKIMATRIX	6,562,360
+ HEURISTIC FILTERING	2,437,108
+ LASER FILTERING	1,250,683
+ divergentmBERT FILTERING	751,792

Table 4.3: WikiMatrix EN-FR corpus statistics.

Development and Test data We use the official development and test splits of the TED corpus (Qi et al., 2018), consisting of 4,320 and 4,866 gold-standard translation pairs, respectively. All models share the same BPE vocabulary. We report average results across runs with three different random seeds.

Preprocessing We use the standard Moses scripts (Koehn et al., 2007) for punctuation normalization, true-casing, and tokenization. We learn 32K BPEs (Sennrich et al., 2016b)

using SentencePiece (Kudo and Richardson, 2018).

Models We use the base Transformer architecture (Vaswani et al., 2017), with embedding size of 512, transformer hidden size of 2,048, 8 attention heads, 6 transformer layers, and dropout of 0.1. Target embeddings are tied with the output layer weights. We train with label smoothing (0.1). We optimize with Adam (Kingma and Ba, 2015) with a batch size of 4,096 tokens and checkpoint models every 1,000 updates. The initial learning rate is 0.0002, and it is reduced by 30% after 4 checkpoints without validation perplexity improvement. We stop training after 20 checkpoints without improvement. We select the best checkpoint based on validation BLEU (Papineni et al., 2002). All models are trained on a single GeForce GTX 1080 GPU.

Evaluation We evaluate the *translation quality* of the resulting translation models with BLEU and METEOR automatic metrics and the uncertainty of their predictions as the average token probabilities given a reference text. Finally, we measure the percentage of *degenerated hypotheses* by checking if a hypothesis falls under odd repetitions not supported by the reference. When measuring repeated n-grams, we exclude punctuation and conjunctions.

4.4 Findings

We explore the impact of fine-grained semantic divergences on different aspects of NMT models. Concretely, we start our exploration by focusing on translation quality results as measured by standard MT automatic metrics in §4.4.1, then we analyze why fine-grained divergences interact with the uncertainty of NMT models' predictions at training and inference time in §4.4.2, and finally, we study how divergences connect to the generation of a specific type of pathologies in §4.4.3.

	BLEU						METEOR					
	0%	10%	20%	50%	70%	100%	0%	10%	20%	50%	70%	100%
REPLACEMENT	30.89 +0.00	31.00 +0.11	30.82 -0.07	30.40 -0.49	29.74 -1.15	27.01 -3.88	33.74 +0.00	33.63 -0.11	33.66 -0.08	33.54 -0.20	33.12 -0.62	31.02 -2.72
DELETION	30.89 +0.00	30.80 -0.09	30.62 -0.27	28.95 -1.94	29.00 -1.89	27.50 -3.39	33.74 +0.00	33.61 -0.13	33.38 -0.36	32.17 -1.57	32.09 -1.65	31.44 -2.30
SUBSTITUTION	30.89 +0.00	30.72 -0.17	30.49 -0.40	25.04 -5.85	26.57 -4.32	25.18 -5.71	33.74 +0.00	33.56 -0.18	33.50 -0.24	29.59 -4.15	31.58 -2.16	30.75 -2.99

Table 4.4: Results for FR→EN translation on the TED test set (means of 3 runs). Bars denote degradation over EQUIVALENTS (i.e., 0%) across different % of corruption. Divergences hurt BLEU and METEOR when they overwhelm the training data. Transformers are particularly sensitive to fine nuances introduced by LEXICAL SUBSTITUTION.

4.4.1 Translation Quality

Table 4.4 presents the impact of semantic divergences on BLEU and METEOR. Corrupting equivalent bitext with fine-grained divergences hurts translation quality across the board. In most cases, the degradation is proportional to the percentage of corrupted training samples. LEXICAL SUBSTITUTION causes the largest degradation for both metrics. The degradation is relatively smaller for METEOR than BLEU, which we attribute to the fact that METEOR allows matches between synonyms when comparing references to hypotheses. SUBTREE DELETION and LEXICAL SUBSTITUTION corruptions lead to significant degradation at $\geq 50\%$ (BLEU; standard deviations across reruns are < 0.4). By contrast, Transformers are more robust to PHRASE REPLACEMENT corruptions, as degradations are only significant after corrupting $\geq 70\%$ (BLEU) of equivalents.

4.4.2 Token Uncertainty

We measure the impact of divergences on model uncertainty at training time and at test time. For the first, we extract the probability of a reference token conditioned on reference

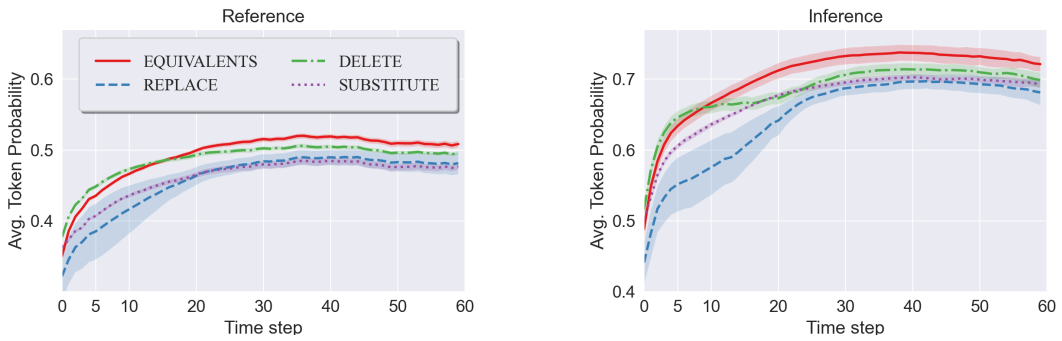


Figure 4.2: Average token probabilities of predictions conditioned on gold references (left) and beam search (5) prefixes (right). Training on fine-grained divergences (100% corruption) increase NMT model's uncertainty.

prefixes at each time step. For the latter, we compute the probability of the token predicted by the model, given its own history of predictions. Figure 4.2 shows that models trained on EQUIVALENTS are more confident in their token-level predictions both at inference and training time. SUBTREE DELETION mismatches affect models' confidence less than other types, while PHRASE REPLACEMENT hurts confidence the most both at inference and at training time. Finally, we observe that differences across divergence types are larger in early decoding steps, while at later steps, they all converge below the EQUIVALENTS.

4.4.3 Degenerated Hypotheses

When models are trained on 50% or more divergent samples, the total length of their hypotheses is longer than the references. Manual analysis on models trained with 100% of divergent samples suggests that this length effect is partially caused by *degenerated* text. Following Holtzman et al. (2019)—who study this phenomenon for unconditional text generation—we define *degenerations* as "output text that is bland, incoherent, or gets stuck in repetitive loops". For instance, "I've never studied sculpture, engineering, and architecture, **and the engineering and architecture"**.

We automatically detect degenerated text in model outputs by checking whether they

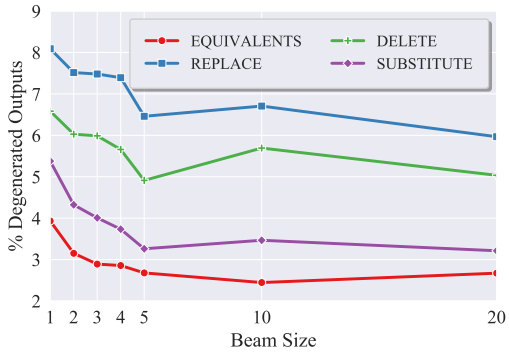


Figure 4.3: % of degenerated outputs as a function of beam size. NMT training on fine-grained divergences (100% corruptions) increases the frequency of degenerated hypotheses across beams.

contain repetitive loops of n-grams that do not appear in the reference.Figure 4.3 shows that exposing NMT to divergences increases the percentage of degenerated outputs. Even with large beams, the models trained on divergent data yield more repetitions than the EQUIVALENTS. Moreover, divergences due to phrasal mismatches (PHRASE REPLACEMENT and SUBTREE DELETION) yield more frequent repetitions than token-level mismatches (LEXICAL SUBSTITUTION). Interestingly, the latter almost matches the frequency of repetitions in EQUIVALENTS with larger beams (≥ 5).

4.5 Summary

Synthetic divergences hurt translation quality, as expected. More surprisingly, our study also reveals that this degradation is partially due to more frequent degenerated outputs and that divergences impact models' confidence in their predictions. Different types of divergences have different effects: LEXICAL SUBSTITUTION causes the largest degradation in translation quality, SUBTREE DELETION and PHRASE REPLACEMENT increase the number of degenerated beam hypotheses, while PHRASE REPLACEMENT also hurts the models' confidence the most. Nevertheless, the impact of divergences on BLEU appears to be smaller

than that of noise (Khayrallah and Koehn, 2018).

Those findings suggest that noise filtering techniques are suboptimal to deal with fine-grained divergences. A natural question that arises is: ***How can we mitigate the negative impact of semantic divergences on*** **NMT*?*** In the next chapter, we aim to answer this question by proposing techniques that are based on two orthogonal approaches: the first incorporates divergent signals in NMT training, while the second focuses on revising divergences in bitexts with the goal of making them more equivalent.

Chapter 5: Mitigating the Impact of Fine-grained Semantic Divergences on Neural Machine Translation

In our third piece of work, we explore different approaches to mitigating the negative impact of fine-grained semantic divergences that is discussed in Section §4. In contrast to the standardized NMT setting that treats all types of divergences (i.e., ranging from noise to fine-grained nuances in bitexts) as either equivalent translations or noisy samples that should be excluded from training, our previous analysis suggests that fine-grained divergences have still the potential to provide *partially* useful training signal. As a result, instead of filtering them out, we propose two interventions to the standard MT pipeline as mitigation strategies: **NMT Intervention** and **Bitext Intervention** (see Figure 5.1).

Our first strategy proposes an **NMT Intervention**: recognizing that divergences exist in the given bitexts, we intervene in the assumption of translation equivalence and aim to model divergences when training NMT systems. The main challenge we face when implementing such interventions is their reliance on a notion of divergence vs. equivalence in bitexts which, as discussed in Section §3, is not available at scale. To overcome this, we draw from our prior work on automatically detecting divergences (Section §3) and propose a divergent-aware NMT framework—DIV-FACTORIZED (§5.1)—that incorporates token-level divergence signals into NMT.

Our second strategy proposes a **Bitext Intervention** based on an orthogonal mitigation direction: instead of altering NMT training to model divergences closely, we aim to automat-

ically re-write them to match the assumption of translation equivalence better. The main challenge we face in this direction is the lack of bilingual supervision for learning the task at hand (i.e., pairs of *divergent-equivalent* bitexts). In our work, we discuss two approaches to overcome this challenge: the first—EQUIV SEMDIV (§5.2)—relies on synthetic translations that selectively replace divergent references under a semantic equivalence condition, and the second—BITEXTEDIT (§5.3)—learns a bitext edit model under noisy supervision based on distance-based mining of imperfect translations.

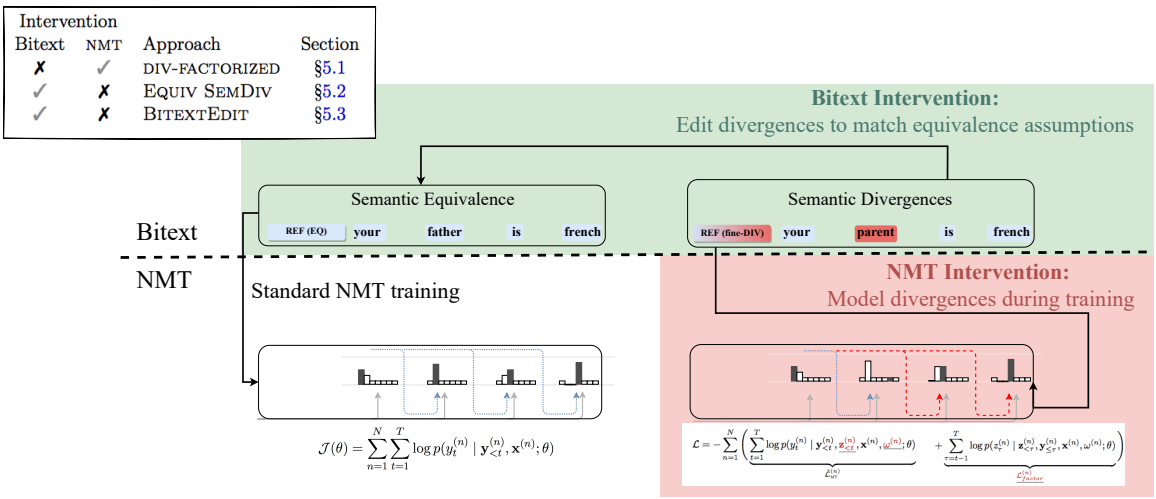


Figure 5.1: Intervention strategies for mitigating the impact of semantic divergences on NMT.

5.1 DIV-FACTORIZED: Factorizing Semantic Divergences for Neural Machine Translation[1]

We introduce a **divergent-aware NMT framework** that incorporates information about which tokens are indicative of semantic divergences between the source and target side of a training sample. Source-side divergence tags are integrated as feature factors (Haddow

[1] **Eleftheria Briakou** & Marine Carpuat. "Beyond Noise: Mitigating the Impact of Fine-grained Semantic Divergences on Neural Machine Translation." *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), 2021.*

and Koehn, 2012; Sennrich and Haddow, 2016; Hoang et al., 2016), while target-side divergence tags form an additional output sequence generated in a multi-task fashion (García-Martínez et al., 2016, 2017). Results on EN↔FR translation show that our approach is a successful **mitigation strategy**: it helps NMT recover from the negative impact of fine-grained divergences on translation quality, with fewer degenerated hypotheses, and more confident and better-calibrated predictions. We make our code publicly available: https://github.com/Elbria/xling-SemDiv-NMT.

5.1.1 Approach

We use **semantic factors** to inform NMT of tokens that are indicative of meaning differences in each sentence pair. We tag divergent source and target tokens in parallel segments as equivalent (EQ) or divergent (DIV) using our mBERT-based token-tagger discussed in Section § 3.3:

SOURCE	TOKENS	*votre père est francais*
	FACTORS	EQ DIV EQ EQ
TARGET	TOKENS	*your parent is french*
	FACTORS	EQ DIV EQ EQ

Source Factors We follow Sennrich and Haddow (2016) who represent the encoder input as a combination of token embeddings and linguistic features. Concretely, we look up separate embedding vectors for tokens and source-side divergent predictions, which are then concatenated. The length of the concatenated vector matches the total embedding size.

Target Factors Target-side divergence tags are an additional output sequence, as in García-Martínez et al. (2016). At each time step, the model produces two distributions: one over the token target vocabulary and one over the target factors. The model is trained to

minimize a divergent-aware loss (Equation 5.1). Terms in red (also underlined) correspond to modifications to the traditional NMT loss. At time step t, the model is rewarded to match the reference target $y_t^{(n)}$, conditioned on the source sequence of tokens ($\mathbf{x}^{(n)}$), the source factors ($\omega^{(n)}$), the token target prefix ($\mathbf{y}_{<t}^{(n)}$), and the target factors prefix ($\mathbf{z}_{<t}^{(n)}$). At the same time (t), the model is rewarded for matching the factored predictions for the previous time step $\tau = t - 1$. The time shift between the two target sequences is introduced so that the model learns to first predict the reference token at τ and then its corresponding EQ vs. DIV label at the same time step. The factored predictions are conditioned again on $\mathbf{x}^{(n)}$, $\omega^{(n)}$, the target factor prefix $\mathbf{z}_{<\tau}^{(n)}$ and the token prefix ($\mathbf{y}_{\leq\tau}^{(n)}$).

$$\mathcal{L} = -\sum_{n=1}^{N} \Bigg(\underbrace{\sum_{t=1}^{T} \log p(y_t^{(n)} \mid \mathbf{y}_{<t}^{(n)}, \underline{\mathbf{z}_{<t}^{(n)}}, \mathbf{x}^{(n)}, \underline{\omega^{(n)}}; \theta)}_{\tilde{\mathcal{L}}_{\text{MT}}^{(n)}} + \underbrace{\sum_{\tau=t-1}^{T} \log p(z_\tau^{(n)} \mid \mathbf{z}_{<\tau}^{(n)}, \mathbf{y}_{\leq\tau}^{(n)}, \mathbf{x}^{(n)}, \omega^{(n)}; \theta)}_{\underline{\mathcal{L}_{factor}^{(n)}}} \Bigg) \tag{5.1}$$

Inference At test time, input tokens are tagged with EQ to encourage the model to predict an equivalent translation. We decode using beam search to predict the translation sequence. The token predictions are conditioned on both the token and the factors prefixes. The factor prefixes are greedily decoded and thus do not participate in beam search.

5.1.2 Experimental Conditions

Divergences We conduct an extensive comparison of models exposed to different amounts of equivalent and divergent WikiMatrix samples. Starting from the pool of examples identified as divergent under our Divergent mBERT model, we rank and select the most fine-grained divergences by thresholding the `bicleaner` score (Ramírez-Sánchez et al., 2020) at 0.5, 0.7 and 0.8.

Models We compare the factored models (**DIV-FACTORIZED**) for incorporating divergent tokens against:

1. **LASER** models are trained on WikiMatrix pairs with a LASER score greater than 1.04 – the noise filtering strategy recommended by Schwenk et al. (2021a). As discussed in Section §3, thresholding LASER might introduce a number of divergent data in the training pool varying from fine to coarse mismatches (Briakou and Carpuat, 2020).
2. **EQUIVALENTS** models are trained on WikiMatrix pairs detected as exact translations under Divergent mBERT;
3. **DIV-AGNOSTIC** models are trained on equivalent and fine-grained divergent data without incorporating information that distinguishes between them;
4. **DIV-TAGGED** models distinguish equivalences from divergences by appending <EQ> vs. <DIV> tags as source-side constraints (Sennrich et al., 2016a).

Models' details Our models are implemented in the `Sockeye2` toolkit (Domhan et al., 2020).[2] We set the size of factor embeddings to 8, the source token embeddings to 504 and target embeddings to 514, yielding equal model sizes across experiments. All other parameters are kept the same across models, as discussed in Section §4.3, except that target embeddings are not tied with output layer weights for factored models.

Other Data & Preprocessing We use the same preprocessing as well as development and test sets as in Section §4.3, except we learn 5K BPEs as in Schwenk et al. (2021a). DIV-FACTORIZED, DIV-AGNOSTIC, and DIV-TAGGED models are compared in controlled setups that use the same training data. We also evaluate out-of-domain on the `khresmoi-summary` test set for the WMT2014 medical translation task (Bojar et al., 2014).

[2] https://github.com/awslabs/sockeye

METHOD		Training size	FR→EN BLEU	FR→EN METEOR	EN→FR BLEU	EN→FR METEOR
LASER		1.25M	31.80 ±0.36	34.00 ±0.17	32.16 ±0.29	56.49 ±0.24
EQUIVALENTS		0.75M	32.88 ±0.07	34.75 ±0.10	33.53 ±0.35	57.38 ±0.28
+DIV	AGNOSTIC	0.93M	32.47 ±0.40	34.56 ±0.20	33.19 ±0.30	57.10 ±0.30
+DIV	TAGGED	0.93M	31.76 ±1.61	34.17 ±0.91	**33.43** ±0.39	**57.55** ±0.27 ↑
+DIV	FACTORIZED	0.93M	32.73 ±0.38	**34.84** ±0.21 ↑	**33.92** ±0.38 ↑	**57.63** ±0.28 ↑
+DIV	AGNOSTIC	1.12M	32.53 ±0.46	34.40 ±0.21	31.47 ±0.61	56.25 ±0.46
+DIV	TAGGED	1.12M	32.38 ±0.40	34.52 ±0.13	**33.35** ±0.17 ↑	**57.33** ±0.14 ↑
+DIV	FACTORIZED	1.12M	32.79 ±0.24	**34.89** ±0.12 ↑	**33.22** ±0.35 ↑	**57.31** ±0.30 ↑
+DIV	AGNOSTIC	1.68M	31.40 ±0.21	33.79 ±0.11	29.53 ±0.39	54.29 ±0.44
+DIV	TAGGED	1.68M	31.97 ±0.26 ↑	34.30 ±0.10 ↑	31.37 ±0.12 ↑	55.87 ±0.18 ↑
+DIV	FACTORIZED	1.68M	32.57 ±0.19 ↑	**34.70** ±0.11 ↑	31.60 ±0.42 ↑	56.10 ±0.22 ↑

Table 5.1: Results for EN↔FR translation on the TED test set (averages and stdev of 3 runs). We underline the top scores among all models and boldface the scores lying within one stdev from EQUIVALENTS. ↑ denotes (one stdev) improvements of DIV-TAGGED and DIV-FACTORIZED over DIV-AGNOSTIC. Factorizing divergences helps NMT recover from the degradation caused by divergences, while it achieves comparable scores to EQUIVALENTS.

Evaluation We evaluate translation quality with BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005).[3,4] We compute Inference Expected Calibration Error (InfECE) as Wang et al. (2020b), which measures the difference in expectation between confidence and accuracy.[5] We measure token-level translation accuracy based on Translation Error Rate (TER) alignments between hypotheses and references.[6] Unless mentioned otherwise, we decode with a beam size of 5.

5.1.3 Results

We discuss the impact of real divergences along the dimensions surfaced by the synthetic data analysis in Section 4, i.e., translation quality, output degeneration, and confidence.

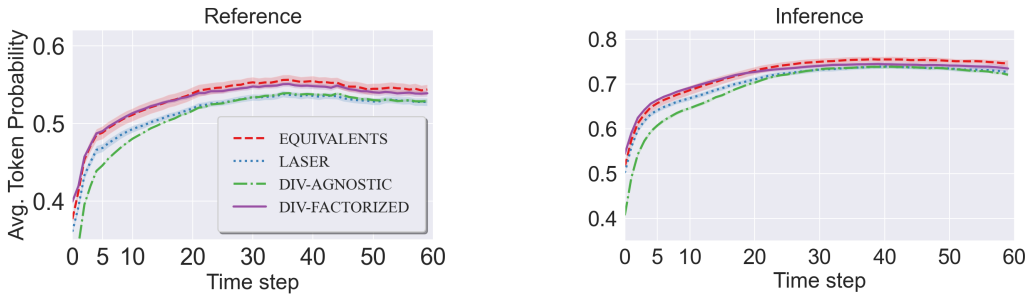
[3] https://github.com/mjpost/sacrebleu
[4] https://www.cs.cmu.edu/~alavie/METEOR/
[5] https://github.com/shuo-git/InfECE
[6] http://www.cs.umd.edu/~snover/tercom/

METHOD	Train. size (M)	FR→EN	EN→FR
LASER	1.25	38.27 ±0.49	39.27 ±0.45
EQUIVALENTS	0.75	39.47 ±0.24	39.63 ±0.52
DIV-AGNOSTIC	0.93	39.45 ±0.50	39.78 ±0.37
	1.12	**40.00** ±0.14	39.20 ±0.50
	1.68	**39.90** ±0.14	38.00 ±0.50
DIV-FACTORIZED	0.93	**40.27** ±0.49	40.13 ±0.46
	1.12	**40.03** ±0.42	**40.30** ±0.29
	1.68	**39.97** ±0.26	39.30 ±0.16

Table 5.2: BLEU scores on the medical domain. We underline top scores and boldface (one stdev) improvements over EQUIVALENTS. Divergences improve translation quality when modeled by DIV-FACTORIZED.

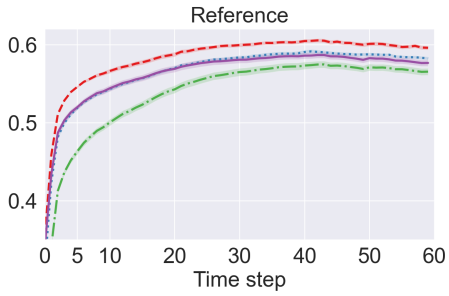
Translation Quality Table 5.1 presents BLEU and METEOR scores across model configurations and data settings on the TED test sets. First, the model trained on EQUIVALENTS represents a very competitive baseline as it performs better or is statistically comparable to all models. This result is in line with prior evidence of Vyas et al. (2018a) who show that filtering out the most divergent pairs in noisy corpora (e.g., OpenSubtitles and CommonCrawl) does not hurt translation quality. Interestingly, the EQUIVALENTS model outperforms LASER across metrics and translation directions, despite the fact that it is exposed to only about half of the training data. Gradually adding divergent data (DIV-AGNOSTIC) hurts translation quality across the board compared to the EQUIVALENTS model. The drops are significantly larger when divergences overwhelm the equivalent translations, which is consistent with our findings on synthetic data.

Second, DIV-FACTORIZED is the most effective mitigation strategy. With segment-level constraints (DIV-TAGGED), models can recover from the degradation caused by divergences (DIV-AGNOSTIC), but not consistently. By contrast, token-level factors (DIV-FACTORIZED) help NMT recover from the impact of divergences across data setups and reach translation quality comparable to that of the EQUIVALENTS model, successfully mitigating the impact of the noisy training signals from divergent samples.

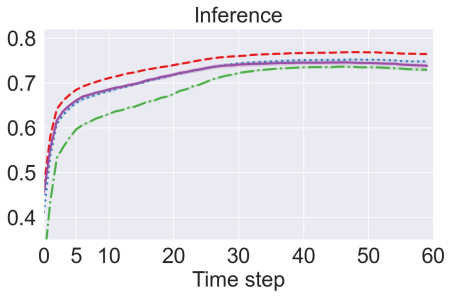


(a) FR→EN

(b) FR→EN



(c) EN→FR



(d) EN→FR

Figure 5.2: Average token probability across time steps on TED test set. DIV-AGNOSTIC yield the least confident predictions (for reference and inference prefixes); DIV-FACTORIZED help recover from this drop (55% divergences).

Third, when translating the out-of-domain test set, DIV-FACTORIZED improves over the EQUIVALENTS model, as presented in Table 5.2. DIV-AGNOSTIC models perform comparably to EQUIVALENTS, while factorizing divergences improves on the latter by $\approx +1$ BLEU, for both directions. Mitigating the impact of divergences is thus important for NMT to benefit from the increased coverage of out-of-domain data provided by the divergent samples.

Uncertainty Figures 5.2a and 5.2c show that the gold-standard references are assigned lower probabilities by the DIV-AGNOSTIC models than all other models, especially in early time steps ($t < 30$). We observe similar drops in confidence based on the probabilities of predicted tokens at inference time (5.2b and 5.2d). This confirms that exposing models to fine-grained semantic divergences hurts their confidence, whether the divergences are

		FR→EN			EN→FR		
METHOD	Train. size	CONF. (↑)	ACC. (↑)	INFECE (↓)	CONF.% (↑)	ACC. (↑)	INFECE (↓)
LASER	1.25M	69.09 ±0.67	62.55 ±0.29	12.34 ±0.38	71.88 ±0.30	60.20 ±0.18	15.10 ±0.12
EQUIVALENTS	0.75M	70.96 ±0.94	63.49 ±0.10	12.37 ±0.24	74.35 ±0.23	61.81 ±0.30	15.09 ±0.18
DIV-AGNOSTIC	0.93M	71.19 ±0.33	63.54 ±0.54	**12.00** ±0.06	73.67 ±0.11	61.44 ±0.22	15.19 ±0.17
DIV-FACTORIZED		**72.16** ±0.10*	**64.29** ±0.44*	**11.81** ±0.04*	74.50 ±0.02*	**62.26** ±0.27*	**14.70** ±0.25*
DIV-AGNOSTIC	1.12M	71.65 ±0.18	61.34 ±0.33	**11.98** ±0.22	71.72 ±0.38	59.29 ±0.48	15.62 ±0.19
DIV-FACTORIZED		71.83 ±0.03	**64.38** ±0.08*	**11.86** ±0.01	74.09 ±0.14*	61.65 ±0.19*	**14.84** ±0.18*
DIV-AGNOSTIC	1.68M	68.01 ±0.37	61.34 ±0.23	12.63 ±0.23	68.38 ±0.25	56.89 ±0.34	16.24 ±0.27
DIV-FACTORIZED		71.01 ±0.39*	63.65 ±0.07*	**11.75** ±0.35*	71.81 ±0.49*	59.78 ±0.39*	14.95 ±0.02*

Table 5.3: Average token confidence, accuracy, and inference calibration results for EN↔FR translation on the TED test set (average and stdev of 3 runs). We underline top scores and boldface (one stdev) improvements over EQUIVALENTS. * denotes (one stdev) improvements of DIV-FACTORIZED over DIV-AGNOSTIC. DIV-FACTORIZED yield more confident and accurate predictions compared to DIV-AGNOSTIC, yielding the smallest calibration errors.

synthetic or not. Furthermore, factorizing divergences helps mitigate the impact of naturally occurring divergences on uncertainty in addition to translation quality.

We conduct a calibration analysis to measure the differences between the confidence (i.e., *probability*) and the correctness (i.e., *accuracy*) of the generated tokens in expectation. Given that deep neural networks are often miscalibrated in the direction of over-estimation (confidence>accuracy) (Guo et al., 2017), we check whether the increased confidence of DIV-FACTORIZED hurts calibration (Table 5.3). DIV-FACTORIZED models are, on average, more confident *and* more accurate than their DIV-AGNOSTIC counterparts. Interestingly, DIV-AGNOSTIC has smaller calibration errors than EQUIVALENTS and LASER models across the board.

Degenerated Hypotheses We check for degenerated outputs across models, data setups (we account for different percentages of divergences in the training data), and different beam sizes (Table 5.4). As with synthetic divergences, we observe that when real divergences overwhelm the training data (55%), degenerated loops are almost twice as frequent for all beam sizes. This phenomenon is consistently mitigated by DIV-FACTORIZED models across

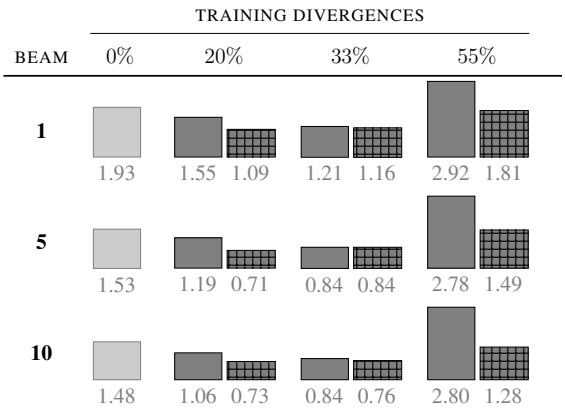


Table 5.4: Percentage of degenerated outputs for FR→EN models exposed to difference percentage of divergent training data (0% corresponds to EQUIVALENTS; dark gray columns correspond to DIV-AGNOSTIC). DIV-FACTORIZED (grid-columns) help recover from degenerations, yielding fewer repetitions across beams.

the board. Furthermore, in some settings (20%, 33%), DIV-FACTORIZED models decrease the amount of degenerated text by half compared to the EQUIVALENTS models. Additionally, LASER models degenerate more frequently than EQUIVALENTS and DIV-FACTORIZED.

5.1.4 Summary

We show that, unlike noisy samples, fine-grained divergences can still provide a useful training signal for NMT when they are modeled via factors. Evaluated on EN↔FR translation tasks, our divergent-aware NMT framework mitigates the negative impact of divergent references on translation quality, improves the confidence and calibration of predictions, and produces degenerated text less frequently. More broadly, this work illustrates how understanding the properties of training data can help build better NMT models. In the following subsections, we explore an orthogonal approach to improving NMT: instead of adapting the standard NMT training paradigm to model divergences more closely, we propose refining them to make them more equivalent.

5.2 EQUIV SEMDIV: Equivalizing Semantic Divergences with Neural Machine Translation[7]

In this section, we explore how synthetic translations can be used to revise potentially imperfect reference translations in mined bitext. Concretely, we present a controlled empirical study comparing the quality of bitext mined from monolingual resources with a synthetic version generated via MT. We focus on the widely used WikiMatrix bitexts for a distant (i.e., EN-EL) and a similar language-pair (i.e., EN-RO) since this corpus contains a significant proportion of erroneous translations discussed in Section 2.1.3. We generate synthetic bitext by translating the original training samples using MT systems trained on the bitext itself and therefore do not inject any additional supervision in the process. We also consider selectively replacing original samples with forward and backward synthetic translations based on our Divergent MBERT classifier, which is also trained without additional supervision (Section §3).

We show that the resulting synthetic bitext improves the quality of the original intrinsically using human assessments of equivalence and extrinsically on bilingual induction (BLI) and MT tasks. We present an extensive analysis of synthetic data properties and of the impact of each step in its generation process. This study brings new insights into the use of synthetic samples in NLP. First, the intrinsic evaluation shows that synthetic translations, in addition to "normalizing" the bitext (Zhou et al., 2019a; Xu et al., 2021), could potentially provide reference translations that are more semantically equivalent to the source than the original ones.

Furthermore, the improved bitext provides more useful signals for BLI tasks and NMT training in two settings (training from scratch; continued training), as confirmed by our

[7] **Eleftheria Briakou** & Marine Carpuat. "Can Synthetic Translations Improve Bitext Quality?" *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022.*

Algorithm 1 Revising Bitext: Given a bitext $\mathcal{D} = (S, T)$, a divergent scorer $\mathbf{R}$, and a margin score t, return revised bitext $\tilde{\mathcal{D}}$

procedure TRAIN($\mathcal{D} = (S, T)$)
    Train $M_{S \to T}$ on $\mathcal{D}$ until convergence
    return $M_{S \to T}$
end procedure
procedure EQUIVALIZE($\mathcal{D} = (S, T)$)
    $M_{S \to T} \leftarrow$ TRAIN($\mathcal{D} = (S, T)$)
    $M_{T \to S} \leftarrow$ TRAIN($\mathcal{D} = (T, S)$)
    $\tilde{\mathcal{D}} \leftarrow \varnothing$
    for $i \in 1,...,|\mathcal{D}|$ **do**
        $(S_i, \hat{T}_i) \leftarrow (S_i, M_{S \to T}(S_i))$
        $(\hat{S}_i, T_i) \leftarrow (M_{T \to S}(T_i), T_i)$
        $d_F \leftarrow \mathbf{R}(S_i, \hat{T}_i) - \mathbf{R}(S_i, T_i)$
        $d_B \leftarrow \mathbf{R}(\hat{S}_i, T_i) - \mathbf{R}(S_i, T_i)$
        if $\max(d_F, d_B) > t$ **then**
            if $\max = d_F$ **then**
                $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \cup \{(S_i, \hat{T}_i)\}$
            else
                $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \cup \{(\hat{S}_i, T_i)\}$
            end if
        else
            $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \cup \{(S_i, T_i)\}$
        end if
    end for
    return $\tilde{\mathcal{D}}$
end procedure

extrinsic evaluations. Finally, ablation analyses that compare different ways to combine synthetic translations show that using *both translation directions* and *filtering using semantic equivalence* is key to improving bitext quality and calls for further exploration of best practices for using synthetic translations in NLP tasks. All bitexts are available at https://github.com/Elbria/xling-SemDiv-Equivalize.

5.2.1 Approach

In this section, we describe the methods and data we use to produce revised bitexts for our empirical study.

Methods for Revising Bitext

We rely on established techniques that can be applied using only the bitext that we seek to revise. First, we train NMT models on the original bitext to translate in both directions. For each original sentence pair, we generate a pool of synthetic translations using NMT and apply a divergence ranking criterion to decide whether and how to replace the original references with a better translation. Algorithm 1 gives an overview of the process, and we describe each step below.

Generating synthetic translations We train NMT models $M_{S\to T}$ and $M_{T\to S}$ on the original bitext to translate in each direction (lines 2-3). For each sentence-pair, they are used to generate two candidates for replacement by forward and backward translation (lines 6-7): $(S_i, M_{S\to T}(S_i))$ and $(M_{T\to S}(T_i), T_i)$. As a result, NMT models translate the exact same data that they are trained on. We thus expect translation quality to be high and that local errors in the original bitext might be corrected by the translation patterns learned by NMT models on the entire corpus.

Selective Replacement We propose to replace an original pair by a candidate *only if* the candidate is predicted to better convey the meaning of the source than the original, which we refer to as the *semantic equivalence condition*. We implement this by ranking the original sample (S_i, T_i), its revision by forward translation $(S_i, M_{S\to T}(S_i))$, and its revision by back-translation $(M_{T\to S}(T_i), T_i)$, according to their degree of semantic equivalence. If none of the synthetic samples score higher than the original, it is not replaced (line 17). Otherwise, the original is replaced by the highest-scoring synthetic sample (lines 10-15). As a result, the cardinality of the bitext remains constant. The semantic equivalence condition (d_F and d_B (lines 8-9)) is implemented using divergentmBERT, a divergent scorer introduced in Section §3 that is trained on synthetic samples generated by perturbations of the original

bitext (e.g., deletions, lexical or phrasal replacements) performed without any bilingual information.

5.2.2 Experimental Conditions

Bitext We evaluate the use of synthetic translations for revising bitext on two language pairs of the WikiMatrix corpus (Schwenk et al., 2021a). WikiMatrix consists of sentence pairs mined from Wikipedia pages using language-agnostic sentence embeddings (LASER) (Artetxe and Schwenk, 2019). Prior work indicates that, as expected, the corpus as a whole comprises many samples that are not exact translations: Kreutzer et al. (2022b) report that for more than half of the audited low-resource language-pairs, mined pairs are on average misaligned; in our prior work we find that 40% of a random sample of the English-French bitext are not semantically equivalent, and include fine-grained meaning differences in addition to alignment noise (Section §3.1). We focus on bitexts with fewer than one million sentence pairs in Greek↔English (EL↔EN, with $750{,}585$ pairs) and Romanian↔English (RO↔EN, with $582{,}134$ pairs), because improving bitext is particularly needed in this data regime. In much higher resource settings, filtering strategies might be sufficient as there might be more high-quality samples overall. In much lower resource settings, the data is likely too noisy or too small to effectively revise bitexts using NMT. We filter out noisy pairs in the training data using `bicleaner` (Ramírez-Sánchez et al., 2020) so that our empirical study excludes the most obvious forms of noise and focuses on the harder case of revising samples that standard preprocessing pipelines consider to be clean.[8]

Preprocessing We use Moses (Koehn et al., 2007) for punctuation normalization, truecasing, and tokenization. We learn 32K BPEs (Sennrich et al., 2016b) per language using

[8] https://github.com/bitextor/bicleaner

`subword-nmt` [9].

NMT Models We use the base Transformer architecture (Vaswani et al., 2017) and include details on the exact architecture and training in Appendix of Briakou and Carpuat (2022a).

Selective Replacement The divergence ranking models are trained using our public implementation of divergentmBERT (Briakou and Carpuat, 2020).[10], following the setup described in Section §3. We use the same margin value for the margin score of Algorithm 1.

5.2.3 Intrinsic Evaluation Results

Human evaluation

We ask 3 bilingual speakers to evaluate the quality of the EN-EL bitexts. Given an original source sentence, they are asked to rank the original target and the candidate target in the order of their equivalence to the source. They are asked "Which sentence conveys the meaning of the source better?", and ties are allowed. A random sample of 100 pairs from forward and backward MT is annotated.

As can be seen in Table 5.5, 60% of ALL synthetic candidates are better translations of the WikiMatrix reference, which confirms the potential of NMT for improving over original translations. Further ablations confirm the benefits of selecting these synthetic candidates with the semantic equivalence condition. When the divergent scorer ranks a candidate higher than the original by a small margin (i.e., $0 \leq d \leq 5$ given $d = R(S_i, M_{S \to T}(T_i)) - R(S_i, T_i)$)), the human evaluation shows that the candidate is actually better than the original only 51% of the times. When using our exact semantic equivalence condition ($d > 5$), candidates are judged as more equivalent than the original 87.5% of the times, and

[9] https://github.com/rsennrich/subword-nmt
[10] https://github.com/Elbria/xling-SemDiv

annotations within this set have a stronger agreement (i.e., 0.688 Kendall's τ). This indicates that the condition $d > 5$ identifies more clear-cut examples of synthetic translations that fix semantic divergences in the original data and can be thus used for selective replacement of imperfect references by better quality translations.

Candidate set	% Equivalized	Kendall's τ
ALL	60.0%	0.321
$d < 0$	26.4%	0.157
$0 \leq d \leq 5$	51.0%	0.234
$d > 5$	$\mathbf{87.5\%}$	$\mathbf{0.688}$

Table 5.5: Human evaluation results for all evaluated pairs and ablation sets for different thresholds on divergent score differences between candidates and originals (i.e., d).

Further inspection of the annotations reveals that most source-target WikiMatrix examples contain fine meaning differences (56%). In those cases, we observe that most of the content between the sentences is shared, but either small segments are mistranslated (e.g., "London" instead of "Athens" in the first example of Table 5.6), or some information is missing from either side of the pair (e.g., "all six" missing from the target side in the third example of Table 5.6). Furthermore, more coarse-grained divergences are found less frequently (12%)—in those cases, we notice that sentences are usually either topically related or structurally similar (e.g., length, syntax) with a few anchor words (e.g., last example in Table 5.6). Finally, 32% of the times the original WikiMatrix pairs are perfect translations of each other.

How do synthetic translations differ from originals?

Figure 5.3 presents the distribution of lexical differences (i.e., computed using LeD—a score that captures lexical differences based on the percentages of tokens that are not found in two sentences (Niu and Carpuat, 2020)) between original and synthetic translations (in EN) for candidates that replace and do not replace the originals.First, we observe that a

[EL]	WIKIMATRIX	Απεβίωσε στην Αθήνα στις 5 Ιουνίου 1979.
	⌊ GLOSS	*He died in Athens on 5 June 1979.*
[EN]	WIKIMATRIX	He died in London on 5 June 1979.
[EN]	SYNTHETIC TRANSLATION	He died in Athens on 5 June 1979.
[EL]	WIKIMATRIX	Ένας από τους οικισμούς που δημιούργησαν ήταν ο Καραβάς.
	⌊ GLOSS	*Karavas was one of the first settlements they created.*
[EN]	WIKIMATRIX	One of the first towns to be created was Vila Barreto.
[EN]	SYNTHETIC TRANSLATION	One of the first settlements to be created was Karavas.
[EL]	WIKIMATRIX	Και οι έξι λέβητες κατασκευάστηκαν από την Waagner-Biro.
	⌊ GLOSS	*All six boilers were manufactured by Waagner-Biro.*
[EN]	WIKIMATRIX	Boilers were supplied by Waagner-Biro.
[EN]	SYNTHETIC TRANSLATION	All six boilers were manufactured by Waagner-Biro.
[EL]	WIKIMATRIX	Το Διδακτικό προσωπικό της Σχολής είναι υψηλού επιπέδου.
	⌊ GLOSS	*The school's teaching staff is of a high level.*
[EN]	WIKIMATRIX	The medical research level of the school is high.
[EN]	SYNTHETIC TRANSLATION	The teaching staff of the school is high.
[EL]	WIKIMATRIX	Ανήκει στο τριπλό αστρικό σύστημα του Άλφα Κενταύρου.
	⌊ GLOSS	*It belongs to the Alpha Centauri triple star system.*
[EN]	WIKIMATRIX	This is the triple alpha process.
[EN]	SYNTHETIC TRANSLATION	It belongs to the triple star system of Alpha Centauri.
[EL]	WIKIMATRIX	Η εμφάνιση τυφώνων είναι σύνηθες φαινόμενο.
	⌊ GLOSS	*The occurrence of hurricanes is a common phenomenon.*
[EN]	WIKIMATRIX	It is extremely rare: There were only 10 known cases in 1998.
[EN]	SYNTHETIC TRANSLATION	The appearance of hurricanes is a common phenomenon.

Table 5.6: Randomly sampled WikiMatrix pairs with synthetic translations that satisfy $d > 5$. Selective replacement successfully revises divergences of different granularities (highlighted segments) in the original references.

substantial amount of synthetic translations that do not replace original references (40%) corresponds to small LED scores (< 0.1), suggesting that the equivalence criterion could fall back to the original sentence not because of the poor quality of candidate references, but rather due to them being already close to the originals. Furthermore, all synthetic translated instances are represented in almost all bins, with fewer instances found on the extreme bins of > 0.7 LED scores. Finally, synthetic translations that replace original references are mostly concentrated within the range $[0.2, 0.6]$ of LeD scores. This indicates that they share lexical content with the original, which further supports the hypothesis that synthetic translations revise fine-grained meaning differences in WikiMatrix in addition to alignment noise.

		PROPERTY	ORIGINAL	REVISED	δ φ
English (EN)	1 :	# Sentences	750,585	750,585	0.0% φ
	2 :	# Tokens	15,244,413	15,239,474	**−0.3%** ≈
	3 :	# Types	358,681	350,224	**−2.4%** ≈
	4 :	Average Length	20.3	20.3	0% φ
	5 :	Average Coverage	0.78	0.83	**+6.0%** ↑
	6 :	# SHE/HER/HERS Pronouns	45,028	43,629	**−3.1%** ≈
	7 :	# HE/HIS/HIM Pronouns	185,356	194,510	**+4.7%** ↑
	8 :	Complexity	63.03	53.61	**−14.9%** ≈
Greek (EL)	9 :	# Sentences	750,585	750,585	0.0% φ
	10 :	# Tokens	15,743,084	15,611,937	**−0.8%** ≈
	11 :	# Types	526,411	519,558	**−1.3%** ≈
	12 :	Average Length	21.0	20.8	**−1.0%** ≈
	13 :	Average Coverage	0.77	0.83	**+7.0%** ↑
	14 :	# Η/ΤΗΣ/ΤΗΝ Pronouns	792,005	776,947	**−1.9%** ≈
	15 :	# Ο/ΤΟΥ/ΤΟΝ Pronouns	799,249	794,275	**−0.6%** ≈
	16 :	Complexity	24.51	17.85	**−27.0%** ≈

Table 5.7: Comparison of original vs. revised bitext for EN-EL. δ gives percentage differences between them.

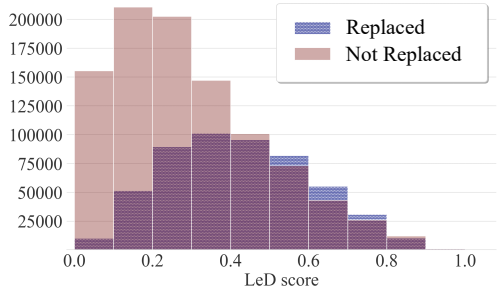


Figure 5.3: LeD differences of original vs. synthetic translations (EL→EN). Replaced candidates share lexical content with the originals.

How does the revised bitext differ from the original?

Table 5.7 presents differences in statistics of the original vs. revised WikiMatrix EN-EL bitexts to shed more light on the impact of selectively using synthetic translation for bitext quality improvement. The refined bitext exhibits higher coverage (i.e., the ratio of source words being aligned by any target words; rows 5 and 13) and smaller complexity (i.e., the

diversity of target word choices given a source word (Zhou et al., 2019a)) compared to the original bitext. Moreover, the use of synthetic translations introduces small decreases in the lexical types covered in the final corpus (i.e., rows 3 and 11), which is expected as the additional coverage in the original corpus might be a result of divergent texts. Those observations are in line with prior work that seeks to characterize the nature of synthetic translations used in other settings, such as knowledge distillation (Zhou et al., 2019a; Xu et al., 2021).

While fixing divergent references contributes to this simplification effect, NMT translations might also reinforce unwanted biases from the original bitext. For instance, the distribution of two grammatical gender pronouns on the English side is a little more imbalanced in the improved bitext than in the original (rows 6-7 and 14-15), likely due to gender bias in NMT (Stanovsky et al., 2019). We limit our analysis to # occurrences for two grammatical gender pronouns. This calls for techniques to mitigate such biases (Saunders and Byrne, 2020; Stafanovičs et al., 2020) for NMT and other downstream tasks.

5.2.4 Extrinsic Evaluation Results

Our previous analysis suggests that selective replacement of divergent references with synthetic translations results in bitext of *improved quality*, with a reduced level of noises and easier word-level mappings between the two languages when compared to the original WikiMatrix corpus. To better understand how those differences impact downstream tasks, we contrast the improved bitext with the original through a series of extrinsic evaluations for EN-EL and EN-RO languages that rely on parallel texts as training samples (§Extrinsic Evaluation Results). First, we focus on the recent state-of-the-art unsupervised BLI approach of Shi et al. (2021) that relies on word alignments of extracted bitexts. Second, we follow the recent bitext quality evaluation frameworks adopted by the "Shared Task on Parallel

Corpus Filtering and Alignment" (Koehn et al., 2020) and built neural machine translation systems from scratch and by continued training on a multilingual pre-trained transformer model. Finally, we conduct extensive ablation experiments to test the impact of using synthetic translations without the semantic equivalence condition and contrast with familiar techniques used by prior work (§Ablation Study).

Experimental Set-Up

BLI The task of BLI aims to induce a bilingual lexicon consisting of word translations in two languages. We experiment with the recently proposed method of Shi et al. (2021) that combines extracted bitext and unsupervised word alignment to perform fully unsupervised induction based on extracted statistics of aligned word pairs. The induced lexicons are evaluated based on MUSE (Lample et al., 2018) consisting of 45,515 and 80,815 dictionary entries for EL-EN and EN-RO, respectively.[11] We extract word alignments using mBERT-based *Simalign*[12] (Jalili Sabet et al., 2020) and statistics based on the implementation of Shi et al. (2021).[13]

MT We experiment with MT tasks following two approaches: (1) training standard transformer seq2seq models from scratch; (2) continued training for mT5 (Xue et al., 2021), a multilingual pre-trained text-to-text transformer. We evaluate translation quality with BLEU (Papineni et al., 2002)[14] on the official development and test splits of the TED corpus (Qi et al., 2018). For (1), we follow the experimental settings described in §5.2.2. For (2), we initialize the weights of the transformer with "mT5-small", which consists of 300M

[11] https://github.com/facebookresearch/MUSE
[12] https://github.com/cisnlp/simalign
[13] https://github.com/facebookresearch/bitext-lexind
[14] https://github.com/mjpost/sacrebleu

parameters,[15]. We use the `simpletransformers` implementation.[16] We fine-tune for up to 5 epochs.

Ablation Settings We compare the NMT models trained on the variants of the synthetic bitext to isolate the impact of replacement criteria and different candidates. For the former, we experiment with the **rejuvenation** approach of Jiao et al. (2020) that replaces original references with forward translated candidates for the 10% least active original samples measured by NMT probability scores. Moreover, we experiment with **forward** and **backtranslation** baselines trained on bitexts that consist solely from target- or source-side candidate sentences (i.e., original references are entirely excluded) and with ablations that consider either forward or backward candidates for the proposed semantic equivalence condition. Finally, we consider two alternatives to the **semantic equivalence** condition based on divergent scores: the **ranking** condition replaces a candidate if it scores higher than the original (i.e., margin with $d = 0$) and the **thresholding** condition adds the additional constraint that candidates should rank higher than a threshold to replace the original pair.

Extrinsic Evaluation Results

BLI Table 5.8 presents results for unsupervised BLI on the MUSE gold-standard dictionaries, for EL-EN and EN-RO. Across languages, the revised bitexts induce better lexicons compared to the original WikiMatrix. Crucially, improvements are reported both in terms of Recall—which connects to the observation that the revised bitext exhibits higher coverage than the original and in terms of Precision—which connects to the noise reduction effect that impacts the extracted word alignments. Additionally, a break-down of the Precision of the induced lexicons binned by the frequency of MUSE source-side entries (i.e., last 3 columns

[15] https://github.com/google-research/multilingual-t5
[16] https://github.com/ThilinaRajapakse/simpletransformers

PAIR	BITEXT	*All* Precision	Recall	F1	OOV rate	*Low*	*Medium* Precision	*High*
EL-EN	Original	76.2	58.1	65.9	**6.7**%	59.4	76.6	81.4
	Revised	**77.6***	**58.6***	**66.8***	7.5%	**60.4***	**78.4***	81.6
EN-RO	Original	89.2	69.4	78.1	**15.8**%	78.6	86.9	87.1
	Revised	**90.8***	**71.3***	**79.8***	16.5%	**80.0***	**87.5***	86.9

Table 5.8: Unsupervised BLI extrinsic evaluation results on MUSE for the entire dataset (*All*) and on subsets binned by frequency (i.e., right-most highlighted columns). Revised bitexts yield statistically significant ($*$) improvements over the original bitexts overall and for low-to-medium frequency dictionary entries.

in Table 5.8) reveals that the improvements come from better induction of low- and medium-frequency words, which we expect are more sensitive to noisy misalignments that result from divergent bitext. Finally, those improvements are reported despite the small increase of the OOV rate in the revised lexicons that results from the decrease in the lexical types covered in it, as mentioned in the analysis under Section §5.2.3.

Furthermore, following the advice of Kementchedjhieva et al. (2019), who raise concerns on BLI evaluations based on gold-standard pre-defined dictionaries, we accompany our evaluation with manual verification to confirm that our conclusions are consistent with those of the automatic evaluation. Concretely, we manually check the *false positives* induced translation pairs from the original vs. the improved bitext. We found that $65/80$ are *false false positives* (due to incompleteness of pre-defined dictionaries) for the improved bitext and $51/80$ for the original. This confirms that the metric improvements we observe are meaningful and suggests that the improved bitext help learn better mappings between source and target words.

MT Table 5.9 presents translation quality (BLEU) on EN↔RO and EN↔EL tasks for MT training from scratch and Figure 5.4 shows translation quality of mT5 continued training across epochs. Across tasks and settings, the revised bitext yields better translation quality

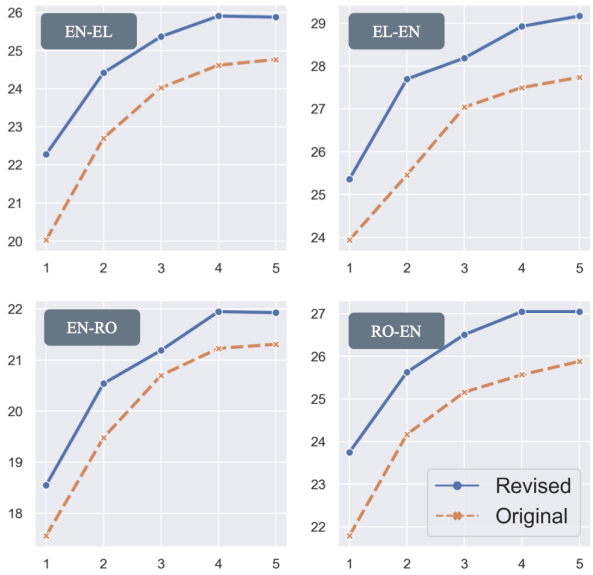


Figure 5.4: BLEU scores across epochs (x-axis) for continued training on mt5. The revised bitext improves translation quality compared to the original for all epochs and translation tasks.

PAIR	ORIGINAL	REVISED
EL→EN	28.15 ±0.13	**29.63 ±0.29**
EN→EL	27.08 ±0.18	**27.89 ±0.05**
RO→EN	23.68 ±0.12	**24.54 ±0.06**
EN→RO	20.65 ±0.10	**20.84 ±0.04**

Table 5.9: BLEU on NMT training from scratch. Revised bitexts improve NMT translation quality.

than the original WikiMatrix data. The consistent improvements we observe across the two settings suggest that the properties of the synthetic translations that replace original samples and bring those improvements are invariant to specific models. Moreover, the magnitude of improvements is larger in the continued training setting compared to training from scratch (e.g., $\sim +0.8$ vs. $\sim +1.5$, for EN→EL; $\sim +0.2$ vs. $\sim +1.5$, for RO→EN). The latter suggests that improvements from using synthetic samples do not only come from the normalization effect (i.e., synthetic samples are easier to model by NMT) but also connect to the reduced noise in the training samples. This further complements our hypothesis

that synthetic translations can improve the quality of imperfect references that should, in principle, yield noisy training signals—and thus impact the resulting quality—of different MT models.

Ablation Study

Table 5.10 compares the translation quality (BLEU) of NMT systems trained on different synthetic translations. By forcing the semantic equivalence condition when deciding whether a synthetic translation replaces an original, we revise 50% of the latter yielding the best results across directions with significant improvements (i.e, increases do not lie within 1 stdev of the original's bitext performance) of $+0.81$ (EN→EL, row 9) and $+1.49$ (EL→EN, row 18) points over the original bitext.

Impact of semantic equivalence condition Table 5.10 shows that naively disregarding the original references and training only on synthetic translations gives mixed results: training on *forward-translated* references only (i.e., row 2) gives small improvements ($+0.36$) over the model trained on WikiMatrix for EN→EL, while it performs comparably to it for EL→EN (i.e., row 11). On the other hand, training on *backward* data only (i.e., row 12) improves BLEU by a small margin ($+0.23$) for MT into EN while it hurts BLEU when translating into EL (i.e., row 3). This indicates that the good quality of the synthetic translations cannot be taken for granted and motivates replacing original pairs under conditions that account for semantic controls.

The latter is further confirmed by results on the rejuvenation baseline: replacing candidates for the 10% of the most inactive WikiMatrix samples results in small and insignificant increases in BLEU when compared to models trained on original WikiMatrix data (i.e., rows 1-4 and 10-13). This indicates that rejuvenation might not be well-suited to lower resource settings than the ones it was originally tested on (Jiao et al., 2020). The rejuvenation

	SELECTIVE REPLACEMENT	DATA TYPES	BLEU	δ	BITEXT STATISTICS $\mathcal{O}$	$\mathcal{F}$	$\mathcal{B}$	VIS.
			EN→EL					
1 :	✗	$\mathcal{O}$	27.08 ±0.18	–	100%	0%	0%	
2 :	✗	$\mathcal{F}$	27.45 ±0.06	+0.36	0%	100%	0%	
3 :	✗	$\mathcal{B}$	26.22 ±0.26	−0.86	0%	0%	100%	
4 :	Rejuvenation	$\mathcal{O}$ $\mathcal{F}$	27.24 ±0.11	+0.16	90%	10%	0%	
5 :	Ranking	$\mathcal{O}$ $\mathcal{F}$	27.21 ±0.43	+0.13	22%	78%	0%	
6 :	Thresholding	$\mathcal{O}$ $\mathcal{F}$	27.56 ±0.11	+0.48	78%	21%	0%	
7 :	Semantic equivalence	$\mathcal{O}$ $\mathcal{F}$	27.64 ±0.22	+0.56	63%	37%	0%	
8 :	Semantic equivalence	$\mathcal{O}$ $\mathcal{B}$	27.61 ±0.09	+0.52	66%	0%	34%	
9 :	Semantic equivalence	$\mathcal{O}$ $\mathcal{F}$ $\mathcal{B}$	**27.89** ±0.05	**+0.81**	50%	23%	27%	
			EL→EN					
10 :	✗	$\mathcal{O}$	28.15 ±0.13	–	100%	0%	0%	
11 :	✗	$\mathcal{F}$	28.16 ±0.17	+0.01	0%	100%	0%	
12 :	✗	$\mathcal{B}$	28.38 ±0.09	+0.23	0%	0%	100%	
13 :	Rejuvenation	$\mathcal{O}$ $\mathcal{F}$	28.27 ±0.12	+0.12	90%	10%	0%	
14 :	Ranking	$\mathcal{O}$ $\mathcal{F}$	28.81 ±0.13	+0.67	26%	74%	0%	
15 :	Thresholding	$\mathcal{O}$ $\mathcal{F}$	28.79 ±0.17	+0.64	81%	19%	0%	
16 :	Semantic equivalence	$\mathcal{O}$ $\mathcal{F}$	29.00 ±0.15	+0.85	66%	34%	0%	
17 :	Semantic equivalence	$\mathcal{O}$ $\mathcal{B}$	29.19 ±0.25	+1.05	63%	0%	37%	
18 :	Semantic equivalence	$\mathcal{O}$ $\mathcal{F}$ $\mathcal{B}$	**29.63** ±0.29	**+1.49**	50%	27%	23%	

Table 5.10: BLEU results (averages of 3 seeds) on EN↔EL NMT. δ denotes average improvements over the original bitext. Bitext statistics give percentage of original ($\mathcal{O}$), forward ($\mathcal{F}$), and backward ($\mathcal{B}$) translated candidates. First column shows the selective replacement condition for candidate replacement (when applicable).

technique might be affected by the decreased NMT quality and calibration in lower resource settings. By contrast, using synthetic translations with semantic control mitigates their impact.

Finally, all three semantic control variants based on divergent scores yield bitexts that improve BLEU compared to the original WikiMatrix (i.e., rows 5-8 and 14-18). Among them, the *margin* condition is the most successful, followed by the *thresholding* variant. The breakdown of training statistics reveals the reason behind their differences: the *thresholding* condition is a more strict constraint as it only allows synthetic candidates to replace the original pairs if they are predicted as exact equivalents, allowing for fewer revisions of divergent pairs in WikiMatrix. By contrast, the condition based on *margin* is a contrastive approach that allows for more revisions of the original data (i.e., a candidate might be a

more fine-grained divergent of the source). The *ranking* criterion is the least successful method—this is expected as the divergence ranker is not trained as a regression model.

Impact of bi-directional candidates Considering both forward ($\mathcal{F}$) and backward ($\mathcal{B}$) translated candidates during selective replacement yields further improvements (0.22-0.44 points) over bitext induced by the semantic equivalence condition with candidates from a single NMT model (i.e., rows 7-9 and 16-18). When forward and backward candidates are considered independently, they replace $34 - 37\%$ of the original pairs; in contrast, when considered together, they replace 50% of original WikiMatrix pairs. As a result, there is no perfect overlap between the original pairs replaced by the forward vs. backward model, which motivates the use of both to revise more divergences in WikiMatrix. This finding raises the question of whether using synthetic translations from both directions might benefit other scenarios, such as knowledge distillation.

5.2.5 Summary

In this section, we explored how synthetic translations can be used to revise bitext, using NMT models trained on the exact same data we seek to revise. Our extensive empirical study surprisingly shows that, even without access to further bilingual data or supervision, this approach improves the quality of the original bitext, especially when synthetic translations are generated in both translation directions and selectively replace the original using a semantic equivalence criterion. Specifically, our intrinsic evaluation showed that synthetic translations are of sufficient quality to improve over the original references, in addition to "normalizing" the bitext as suggested by prior work and corpus level statistics (Zhou et al., 2019a; Xu et al., 2021). Extrinsic evaluations further show that the replaced synthetic translations provide more useful signals for BLI tasks and NMT training in two settings (i.e., training from scratch and continued training).

These findings provide a foundation for further exploration of the use of synthetic bitext. Our empirical study focused on language pairs and datasets where revising bitexts is the most needed and most likely to be useful: the resources available for these languages are not so large that mined bitext can simply be ignored or filtered with simple heuristics, yet there is enough data to build NMT systems of reasonable quality (i.e., $\sim$ 600K segments for EN-RO, and $\sim$ 750K for EN-EL). While in principle, selective replacement of divergent references with synthetic translations should port to high-resource settings, where NMT is as good or better than for the languages considered in this work, other techniques are likely needed in low-resource settings where NMT quality is too low to provide reliable candidate translations. In the following section, we introduce an editing-based model for bitext quality improvement that utilizes divergent bitext more effectively in more heterogeneous data scenarios.

5.3 BITEXTEDIT: Automatic Bitext Editing for Improved Bitext Quality[17]

In this section, we propose an *editing approach to bitext quality improvement* that aims to make use of as much of the signal from potentially imperfect mined bitext as possible. Our model takes as input a bitext (i.e., $(\mathbf{x_f}, \mathbf{x_e})$) and edits one of the two sentences to generate a refined version of the original (i.e., $\mathbf{x'_f}$ or $\mathbf{x'_e}$) as necessary. By framing the problem as a bitext editing (BITEXTEDIT) task, we can perform a wide range of operations from *copying* good-quality bitext to *partial editing* of small meaning mismatches and *translating* from scratch incorrect references. Unlike our previous work in Section 5.2 that focused on WikiMatrix bitexts mined from Wikipedia, here we switch to CCMatrix (Schwenk et al., 2021b), a corpus mined from CommonCrawl, which is 50 time larger than WikiMatrix,

[17] **Eleftheria Briakou**, Sida I Wang, Luke Zettlemoyer, Marjan Ghazvininejad. "BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation". *In Proceedings of the Findings of the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL Findings), 2022.*

including several low-resource languages. Following previous extrinsic evaluations of bitext quality (Koehn et al., 2019, 2020; Schwenk et al., 2021b,a), we compare NMT models trained on the original and revised versions of mined bitexts. Concretely, we report consistent improvements in translation quality for 10 low-resource NMT translation tasks: EN↔OC, IT↔OC, EN↔BE, EN↔MR, and EN↔SW, while in most cases, we even improve upon a competitive translation-based baseline. Crucially, BITEXTEDIT yields from ~ 4, up to ~ 8 BLEU point improvements in the more data-scarce settings (i.e., EN-OC, IT-OC). Additionally, our quantitative and qualitative analysis indicates that BITEXTEDIT improves bitext quality in higher-resource settings with lighter editing that targets more fine-grained meaning differences. We make our code available at: https://github.com/Elbria/BitextEdit.

5.3.1 Approach

We frame bitext refinement as an editing task (i.e., BITEXTEDIT) that takes two *input sentences*: a sentence $\mathbf{x_f}$ in language $\mathbf{f}$ and a sentence $\mathbf{x_e}$ in language $\mathbf{e}$, and aims at editing *one* of them (i.e., it *outputs* $\mathbf{x'_f}$ or $\mathbf{x'_e}$) with the goal of yielding a more equivalent translation pair (i.e., <$\mathbf{x_f}$, $\mathbf{x'_e}$> or <$\mathbf{x'_f}$, $\mathbf{x_e}$>). Below, we describe the bitext refinement model (§Bitext Editing) and the curation of data needed to train our model based on bitext mining (§Bitext Mining).

Bitext Editing

Architecture Our bitext editing model is a transformer seq2seq architecture. Each bitext ($\mathbf{x_f}$, $\mathbf{x_e}$) is encoded via adding position embeddings that are reset for each input sentence to facilitate their alignment (Conneau and Lample, 2019) and two language embeddings, initialized at random, to indicate the two languages for the editing model. The decoder generates autoregressively a refined version of $\mathbf{x_f}$ or $\mathbf{x_e}$, where the first generated token

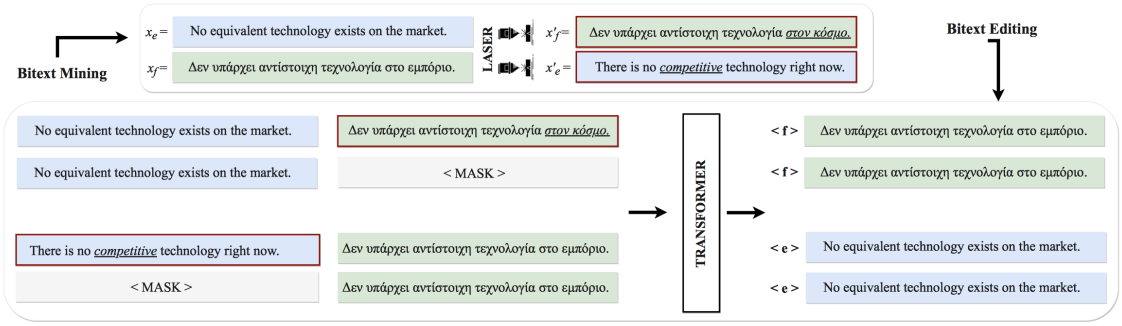


Figure 5.5: BITEXTEDIT training strategy: Our multi-task model is trained using synthetic supervision from mined bitexts. Starting from an original bitext (x_e, x_f), we mine imperfect translations x'_f and x'_e for each reference using LASER (Bitext Mining). A sequence-to-sequence Transformer model is trained to *translate* and *reconstruct* the original references given synthetically extracted bitexts representing imperfect translations (Bitext Editing).

indicates which of the two input sentences is edited, as described below.

Learning As presented in Figure 5.5, during training, we optimize the multi-task loss presented in Equation 5.2, which has two components. The first represents a *edit-based reconstruction loss* (i.e., $\mathcal{L}_{\text{EDIT}}$) that reconstructs one of the two sentences, e.g., $\mathbf{x_f}$ started from a noised version of the original bitexts e.g., $\mathbf{x'_f}$ and $\mathbf{x_e}$. We make this loss bi-directional by adding a symmetrical loss that reconstructs $\mathbf{x_e}$ from $\mathbf{x_f}$ and $\mathbf{x'_e}$, respectively. The second component is implemented as a bi-directional *translation loss* (i.e., $\mathcal{L}_{\text{MT}}$) via masking the inputs of the target translation directions (e.g., generate $\mathbf{x_e}$ given $\mathbf{x_f}$ and <MASK>). Finally, in both losses, a language identification symbol (i.e., <f> or <e>) is used as the initial token to predict the language of the output text.

$$\mathcal{L} = \sum_{(\mathbf{x_f}, \mathbf{x_e})} \Bigg(\underbrace{\text{logp}\Big([\text{<e>}\ \mathbf{x_e}] \mid (\mathbf{x_f}, \mathbf{x'_e})\Big) + \text{logp}\Big([\text{<f>}\ \mathbf{x_f}] \mid (\mathbf{x'_f}, \mathbf{x_e})\Big)}_{\mathcal{L}_{\text{EDIT}}} + \underbrace{\text{logp}\Big([\text{<e>}\ \mathbf{x_e}] \mid (\mathbf{x_f}, \text{<MASK>})\Big) + \text{logp}\Big([\text{<f>}\ \mathbf{x_f}] \mid (\text{<MASK>}, \mathbf{x_e})\Big)}_{\mathcal{L}_{\text{MT}}} \Bigg) \tag{5.2}$$

Inference At test time, our model takes as input a possibly imperfect bitext and edits one of the reference translations while first generating the language identification token. The latter is used to infer which of the two reference translations gets revised. Finally, we pair the edited output sequence with the original input that does not get revised, yielding a refined bitext.

Bitext Mining

Our model requires access to $\mathbf{x}'_{\mathbf{f}}$ and $\mathbf{x}'_{\mathbf{e}}$ training instances that are treated as noised versions of $\mathbf{x}_{\mathbf{f}}$ and $\mathbf{x}_{\mathbf{e}}$, respectively. Since our goal is to develop a model that can refine mismatches found in mined bitexts at inference time, we want our noised training instances to share similar properties with the mined ones (e.g., fluent text in the target language, possibly imperfect translations of the source text). To this direction, we take a distance-based mining approach to construct the noised samples similar to Schwenk (2018).[18] Concretely, given the mined bitext $(\mathbf{x}_{\mathbf{f}}, \mathbf{x}_{\mathbf{e}})$ and two pools of monolingual sentences $\mathcal{F}$ and $\mathcal{E}$, in language $\mathbf{f}$ and $\mathbf{e}$, we extract $\mathbf{x}'_{\mathbf{f}}$ and $\mathbf{x}'_{\mathbf{e}}$ as follows:

$$\begin{aligned} \mathbf{x}'_{\mathbf{f}} &= \text{argmax}_{\mathbf{z}\in\mathcal{F}} \ \cos(\text{LASER}(\mathbf{z}), \text{LASER}(\mathbf{x}_{\mathbf{e}})) \\ \mathbf{x}'_{\mathbf{e}} &= \text{argmax}_{\mathbf{z}\in\mathcal{E}} \ \cos(\text{LASER}(\mathbf{x}_{\mathbf{f}}), \text{LASER}(\mathbf{z})) \end{aligned} \tag{5.3}$$

where LASER (Artetxe and Schwenk, 2019) represents a multilingual encoder used to extract sentence embeddings for each sentence, while the most similar sentence is returned based on nearest neighbor retrieval. Furthermore, this formula is extended to retrieve top k sentences while we also allow mining of the original CCMatrix translations. The latter happens to expose the model to good translations at training time, which should not be edited.

[18] Note that unlike Artetxe and Schwenk (2019) we do not use a margin score on the *normalized* cosine distance of sentence-pairs to keep the computation cost low and encourage mining of more diverse imperfect translations.

5.3.2 Experimental Conditions

Bitexts We focus on CCMatrix data for two main reasons: a) it constitutes the only large-scale available resource for a lot of low-resource language pairs, and b) recent efforts of auditing this corpus raise concerns regarding the quality of mined bitext of low-resource pairs. CCMatrix is mined using LASER embeddings following the "max-strategy" approach: a margin score is computed for all monolingual sentences in two languages, then the union of forward and backward candidates is built, and pairs that score above a pre-defined threshold are treated as translations. Schwenk et al. (2021b) set the threshold globally for all languages at 1.06.

Our primary goal is to explore whether bitexts that are typically discarded by filtering can be refined by our model and thus benefit low-resource NMT. For this purpose, we define two pools of CCMatrix data: Pool A corresponds to CCMatrix data with LASER scores greater than 1.06, while Pool B contains bitexts with scores lower than 1.06 and greater than 1.05. The latter threshold is primarily chosen since CCMatrix bitexts is only available above this value. Editing bitexts with even smaller scores is an interesting area for future work.

Training data BITEXTEDIT models are trained based on procedures described in Section §5.3.1, where we use Pool A to seed the generation of noised training samples $\mathbf{x}'_{\mathbf{f}}$ and $\mathbf{x}'_{\mathbf{e}}$. We mine k samples $\mathbf{x}'_{\mathbf{f}}$ for each $\mathbf{x}_{\mathbf{e}}$ and k samples $\mathbf{x}'_{\mathbf{e}}$ for each $\mathbf{x}_{\mathbf{f}}$, respectively. We set k to 4.

Language-pairs We experiment with the following languages: English-Occitan (EN-OC), Italian-Occitan (IT-OC), English-Belarusian (EN-BE), English-Marathi (EN-MR), and English-Swahili (EN-SW). The 5 language pairs are chosen to include diverse low-resource pairs, which differ either in training data size or language similarity. Table 5.11 summarizes the data statistics.

ISO	PAIR	SCRIPTS	Pool A	Pool B
EN-OC	English-Occitan	Latin-Latin	0.2M	0.1M
IT-OC	Italian-Occitan	Latin-Latin	0.3M	0.1M
EN-BE	English-Belarusian	Latin-Cyrillic	0.7M	1.1M
EN-MR	English-Marathi	Latin-Devanagari	1.5M	2.1M
EN-SW	English-Swahili	Latin-Latin	1.7M	0.9M

Table 5.11: Statistics of CCMatrix bitexts.

Comparisons We run several extrinsic evaluations using NMT trained on different versions of CCMatrix data. First, we train NMT models on two versions of original CCMatrix data: Pool A (Schwenk et al., 2021b) and Pool $A \cup B$. Second, we aim to revise Pool B via a) a translation-based approach that revisits the source side of the bitexts via back-translating their target side with a model trained on the original CCMatrix, (i.e., $b(.)$) and b) via editing either the source or the target side of it using our proposed approach (i.e., $r(.)$).

Model details Our models are implemented on top of `fairseq` (Ott et al., 2019).[19] We use the same Transformer architecture as in Schwenk et al. (2021b), with embedding size 512, 4,096 transformer hidden size, 8 attention heads, 6 transformer layers, and dropout 0.4. We train with 0.2 label smoothing and Adam optimizer with a batch size of 4,000 tokens per GPU. We train for 100 epochs and select the best checkpoint based on validation perplexity. We report single-run results.

Data Preprocessing We use the standard Moses scripts (Koehn et al., 2007) for tokenization of EN, OC, IT, BE and SW and the Indic NLP library[20] for MR. For each language-pair, we learn 60K BPEs using `subword-nmt` (Sennrich et al., 2016b).[21]

[19] https://github.com/pytorch/fairseq
[20] https://anoopkunchukuttan.github.io/indic_nlp_library/
[21] https://github.com/rsennrich/subword-nmt

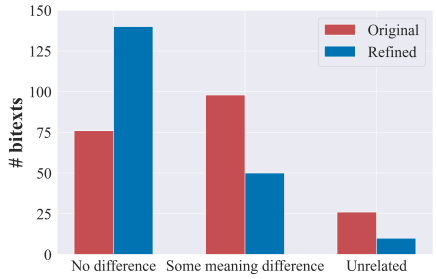


Figure 5.6: Number of bitexts manually rated as perfect translations (i.e., No difference), partial translations (i.e., some meaning difference), and wrong translations (i.e., unrelated) for a random sample of original vs. refined CCMatrix EN-EL data.

Evaluation We evaluate our models on the *devtest* of `flores` (Guzmán et al., 2019). We report `spm-bleu`[22] on the detokenized version of our outputs.

5.3.3 Intrinsic Evaluation Results

We conduct a qualitative study to confirm that BITEXTEDIT improves the quality of CCMatrix. One of the authors manually evaluates a random sample of 200 EN-EL sentence pairs where we compare the original bitexts against the refined ones. Here, we present results on bitext refinement models that use 0.5M PoolA samples. Manual inspection of refined outputs of models trained on larger pools showed similar performance. As shown in Figure 5.6, our models perform edits that refine meaning mismatches found in the original CCMatrix data. While only $\sim 38\%$ of the original samples contain parallel texts that are perfect translations of each other, the revised sample contains $\sim 70\%$ perfect translations. Finally—apart from evaluating meaning differences—we also rate the fluency of the edited translations. We find that our model does not suffer from major fluency issues with 84.5% of their outputs rated as *flawless* and 15.5% as *good*. Table 5.12 presents example outputs of our BITEXTEDIT approach for English-Greek.

[22] https://github.com/facebookresearch/flores

→	**[EN]** CCMATRIX	After that time the whole group would talk for 5 minutes.
	[EL] CCMATRIX	Αργότερα, η ομάδα μελέτης ζήτησε από όλους να διαλογιστούν για πέντε λεπτά.
	⌊ GLOSS	*Later, the study group asked everyone to meditate for 5 minutes.*
	[EN] BITEXTEDIT	Later, the study group asked everyone to meditate for five minutes.
→	**[EN]** CCMATRIX	We should, however, always be striving to live a sustainable and kind life.
	[EL] CCMATRIX	Πάντα πρέπει να παλεύουμε για δίκαιη και βιώσιμη ειρήνη.
	⌊ GLOSS	*We must always fight for a just and lasting peace.*
	[EN] BITEXTEDIT	We must always fight for just and sustainable peace.
	[EN] CCMATRIX	"The western influence came from film and television", he later explained.
→	**[EN]** CCMATRIX	«Η λογοκρισία εντείνεται όλο και περισσότερο στον κινηματογράφο και την τηλεόραση», εξήγησε ο ίδιος.
	⌊ GLOSS	*"Censorship is intensifying in cinema and television", he explained.*
	[EL] BITEXTEDIT	«Η δυτική επιρροή ήρθε από την ταινία και την τηλεόραση» , εξήγησε αργότερα.
	⌊ GLOSS	*"The western influence came from form and television", as their later explained.*
	[EN] CCMATRIX	I could work with a hospital specialist as a clinical assistant (as I have done).
→	**[EL]** CCMATRIX	Δούλευε ώς βοηθός ερευνητή παράλληλα με το διδακτορικό (όπως και εγώ)
	⌊ GLOSS	*They were working as an assistant researcher in parallel with their doctorate (as I have done).*
	[EL] BITEXTEDIT	Θα μπορούσα να δουλέψω με έναν ειδικό στο νοσοκομείο ως κλινικός βοηθός (όπως έχω κάνει).
	⌊ GLOSS	*I could work with a hospital specialist as a clinical assistant (as I have done).*

Table 5.12: Examples of CCMatrix bitexts along with refined sides generated by BITEXTEDIT. → denotes the side ([EL] or [EN]) that the model edits, while highlighted segments indicate the meaning mismatches in the original CCMatrix sentence that gets edited. Greek sentences are glossed to help understanding their meaning.

5.3.4 Extrinsic Evaluation Results

We present our main results on MT tasks that compare BITEXTEDIT with baselines (§Main Results) and then turn into analysis experiments that explore how our model performs in different data regimes (§Analysis).

Main Results

Bitext filtering revisited We first provide empirical evidence that bitext filtering might be a suboptimal solution to low-resource NMT. Table 5.13 shows that filtering out sentence pairs that score below the predefined threshold of 1.06 (i.e., Filtering) surprisingly hurts translation quality in almost all translation tasks (rows 2 vs. 1 and 8 vs. 7). This result is likely because the threshold was optimized for specific language pairs and the fact that—under low-resource regimes—increasing the amounts of *possibly imperfect* translation data

			EN→OC		IT→OC		EN→BE		EN→MR		EN→SW	
1 :	CCMatrix	$A \cup B$	20.5		11.5		11.0		12.2		38.1	
2 :	Filtering	A	18.1	−2.4	11.7	+0.2	9.8	−0.2	12.2	0.0	37.6	−0.5
3 :	Translation-based	$b(A \cup B)$	20.8	+0.3	17.0	+5.5	12.3	+1.3	15.5	+3.2	37.6	−0.5
4 :	BITEXTEDIT	$r(A \cup B)$	25.4	+4.9	19.8	+8.3	12.8	+1.7	**15.8**	+3.6	37.8	−0.3
5 :	Translation-based	$A \cup b(B)$	23.0	+2.5	17.0	+5.5	12.1	+1.1	15.4	+3.2	**38.8**	+0.7
6 :	BITEXTEDIT	$A \cup r(B)$	**26.0**	+5.5	**19.9**	+8.4	**13.0**	+2.0	15.3	+3.1	38.3	+0.2
			OC→EN		OC→IT		BE→EN		MR→EN		SW→EN	
7 :	CCMatrix	$A \cup B$	24.3		11.6		9.8		13.0		34.8	
8 :	Filtering	A	17.8	−6.5	11.1	−0.5	7.8	−2.0	11.3	−1.07	34.8	0.0
9 :	Translation-based	$b(A \cup B)$	26.6	+2.3	17.3	+5.7	9.9	+0.1	13.6	+0.6	33.8	−1.0
10 :	BITEXTEDIT	$r(A \cup B)$	28.2	+3.9	**18.5**	+6.9	10.7	+0.9	16.4	+3.4	35.8	+1.1
11 :	Translation-based	$A \cup b(B)$	27.7	+3.4	15.6	+4.0	9.6	−0.2	15.1	+2.1	**36.8**	+2.0
12 :	BITEXTEDIT	$A \cup r(B)$	**28.7**	+4.4	18.3	+6.7	**10.8**	+1.0	**16.7**	+3.7	36.2	+1.8

Table 5.13: Results on NMT tasks for models trained on different versions of CCMatrix. For each task the first column denotes spm-BLEU; the second columns (highlighted scores) give the difference of each row with the original CCMatrix. Models trained on the refined bitexts improve NMT for low-resource language-pairs.

might still benefit NMT. Furthermore, this experiment gives us insights into the quality of the training data our bitext editing model uses: for IT-OC, BE-EN, and EN-MR we expect Pool A to provide more noisy training signals (as BLEU scores of NMT models trained on it are ~ 11), compared to EN-OC and EN-SW where the quality of the given bitext is expected to be significantly better (BLEU scores ~ 18 and ~ 37, respectively).

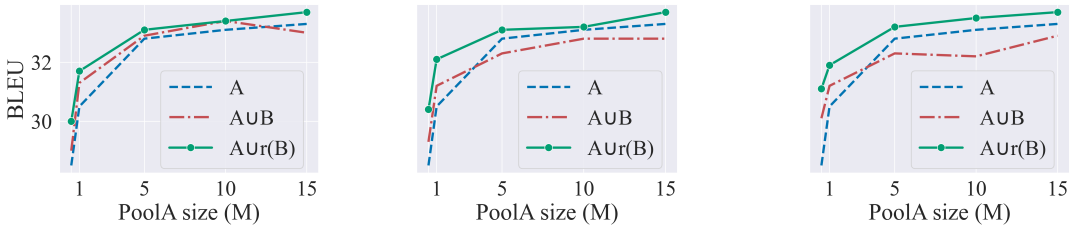
Editing Pool B Applying BITEXTEDIT to edit erroneous translations in Pool B (i.e., $A \cup r(B)$) improves the quality of NMT systems over the ones trained on the original CCMatrix corpus (rows 6 vs. 1 and 12 vs. 7). Among the language pairs considered, the largest improvements are reported for IT-OC translation tasks (i.e., $+8.4/+6.7$), followed by EN-OC (i.e., $+5.5/+4.4$). The magnitude of improvements might be explained by the relatedness of the two languages, which facilitates editing with simpler operations (e.g., copying instead of translating).

Our approach also brings significant improvements over the original data for distant language pairs written in different scripts, despite being trained on more noisy data, as discussed above. For example, we see improvements $+2.0/+1.0$ for EN-BE and $+3.1/+3.7$

for EN-MR. On the other hand, improvements on EN-SW are smaller (i.e., $+0.5/+1.8$). This is expected given the high BLEU scores that the original CCMatrix data yields.

Comparison with Translation-based Baseline Since Pool B bitexts are typically filtered out from the pool of NMT training instances, one reasonable way of incorporating them in NMT training is via treating them as monolingual samples. We experiment with a translation model based on back-translation, which constitutes the most popular approach to employ data augmentation for NMT. Comparing NMT models trained on CCMatrix augmented with back-translated Pool B against our revised Pool B version (i.e., rows 5 vs. 6 and 11 vs. 12) shows that editing outperforms the translation-based model for $7/10$ tasks, while it yields comparable results to it for the rest 3.

Editing Pool A and Pool B Since the editing framework gives us the potential to generalize all types of operations that *might* be needed to refine bitexts, it is also important that it does not perform *overediting* (i.e., editing already good quality bitexts). For this reason, we also attempt to revise the entire CCMatrix corpus (i.e., $r(A \cup B)$) using our bitext refinement models (i.e., rows 4 and 9). To better understand the importance of performing conservative editing on good quality bitexts, we also compare against the translation-based baseline (i.e., $b(A \cup B)$ in rows 3 and 9). First, we observe that our approach yields consistently significant improvements over CCMatrix with the exception of EN$\rightarrow$SW where it performs comparably to it. Second, for most tasks, the improvements are comparable to those reported when revising only Pool B, while it is consistently better than the translation-based model. It, overall, provides a universal method that works well in every case.



(a) $|B| = |A|/2$ (b) $|B| = |A|$ (c) $|B| = 2|A|$

Figure 5.7: Translation quality (i.e., BLEU) of EN→EL NMT models trained on different amounts of Pool A and Pool B data (i.e., $|A|$ given by x-axis). Across settings, bitext refinement (i.e., $A \cup r(B)$) performs better or comparably to training on the original CCMatrix (i.e., $A \cup B$) or its filtered version (i.e., A).

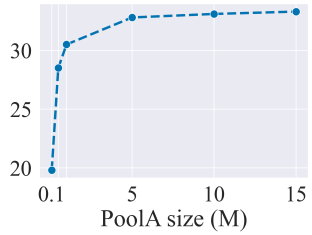


Figure 5.8: BLEU for EN→EL NMT trained on varying size of CCMatrix data (Pool A).

Analysis

Scaling-up BITEXTEDIT First, we examine how models trained only on good quality data (Figure 5.8) behave as we vary their quantity. We experiment with English-Greek EN-EL CCMatrix bitexts and simulate various resource settings via downsampling. In *low-resource* settings (i.e., $|A| < 1$M), translation quality exhibits rapid improvements, with an increase from 100K to 500K training samples boosting BLEU, by approximately 10 points. In *medium-resource* scenarios (i.e., $1 <$M$|A| < 5$M), a proportional increase in the quantity of good quality bitexts yields smaller—yet, significant—translation quality improvements (i.e., moving from 1M to 5M bitexts yields $+2$ BLEU). Finally, in *high-resource* settings (i.e., $|A| > 5$), translation quality reaches a saturation point, with BLEU increases being small and insignificant (i.e., $\sim +0.2$) as we move from 10M to 15M training samples.

Second, we present a controlled analysis experiment on how bitext refinement impacts the

translation quality of NMT systems under different resource settings (Figure 5.7). Concretely, starting from a high resource language-pair in CCMatrix (here, EN-EL), we sample good and poor quality bitexts (i.e., A and B, respectively) representing low- to high-data scenarios (e.g., 500K up to 15M sentence-pairs). Then, we train EN→EL NMT systems on $A \cup B$ while varying their distribution to represent three settings: (a) good quality bitexts overwhelm the training data (i.e., $|B| = |A|/2$), (b) good and poor quality bitext are equally represented (i.e., $|B| = |A|$), and (c) poor quality bitexts overwhelm the training data (i.e., $|B| = 2|A|/$). Across distribution conditions, adding imperfect translations (i.e., B) to the original good-quality data yields improvements for low-to-medium resource settings (i.e., $|A| < 5$). These results complement the earlier observations of §5.3.4 that question the appropriateness of a filtering framework in settings where data is scarce. On the other hand, when moving to high resource scenarios, the additional signal that results from imperfect references can have either insignificant (i.e., Figure 5.7a) or negative impact (i.e., Figures 5.7b and 5.7c) on translation quality. The latter depends on whether the good quality data is underrepresented in the training samples.

Third, starting from good quality bitexts of varying sizes, we train separate bitext refinement models and edit the corresponding poor quality samples (i.e., $r(.)$) defined earlier. Across the board, NMT models that are trained on $A \cup r(B)$ yield the best translation quality results compared to both filtering and training on the original CCMatrix. However, we observe that the magnitude of the improvements depends on the data settings. Concretely, bitext refinement yields significant improvements on low-to-medium resource settings (i.e., $\sim +2$ BLUE points). On the other hand, in high-resource scenarios, bitext refinement helps mitigate the negative impact of overwhelming poor-quality instances and performs comparably to filtering. The latter suggests that our refinement strategy *improves bitexts quality across low- to high-resource settings*.

		C	S	D	I	C	S	D	I
CORPUS	EDITED SENT.	ALL (%)				ALL \ COPIES (%)			
Tatoeba	29.80%	97.47	1.88	0.29	0.34	86.38	10.16	1.56	1.88
OpenSubtitles	65.63%	90.46	5.53	1.27	2.73	74.51	14.79	3.39	7.29
ParaCrawl	88.11%	96.30	2.25	0.39	1.04	85.42	8.89	1.55	4.12

Table 5.14: TER statistics for bitext refinement of random samples of EN-EL OPUS bitexts. Second column gives the % of bitexts that get at least one edit operation; the last two columns present the percentage of correct (**C**), substituted (**S**), deleted (**D**), and inserted (**I**) tokens for all the bitexts (i.e., **ALL**) and the subset of bitexts that receive revisions compared to the original (i.e., **ALL**\ **COPIES**).

PAIR	SRC	TGT	BOTH
EN-OC	34.06%	66.58%	67.48%
IT-OC	34.76%	41.11%	75.78%
EN-MR	58.35%	19.90%	68.07%
BE-EN	21.01%	28.06%	49.06%
EN-SW	14.52%	21.05%	35.57%

Table 5.15: Percentage of sentences with at least one edit operation compared to the original for: source-side (SRC), target-side (TGT), and both sides (BOTH).

Percentage of edited bitexts Table 5.15 presents coarse statistics on the percentage of refined bitexts that exhibit at least one edit compared to the original ones. First, we observe that the percentage of edited bitexts varies across the languages-pairs studied. This reflects the varying quality of PoolB samples in different languages and also connects to the varying magnitude of improvements we show in Table 5.13. The biggest improvements are given for IT-OC, where $\sim 76\%$ of the bitexts are edited by our refinement models. On the other hand, the smallest improvements are found for EN-SW, with only $\sim 36\%$ of its bitext being revised, probably due to the already good quality of the initial CCMatrix sentence pairs.

Editing EN-EL OPUS corpora Broadly speaking, a good bitext refinement model should be able to rewrite bitext in a way that improves potential errors in the original references. At the same time, though, it should avoid over-editing (i.e., avoid editing an already good translation pair). We perform a quantitative analysis on EN-EL corpora from OPUS that vary

in their quality and extract Translation Error Rate (TER) label (Snover et al., 2006) token-level statistics to study both the *frequency* and the *types* of edits that our bitext refinement models perform. Table 5.14 presents results on random samples ($\sim$ 100K) of three popular corpora: (a) the Tatoeba corpus (Tiedemann, 2020) consisting of human translations, (b) the OpenSubtitles corpus (Lison and Tiedemann, 2016) consisting of sentence-aligned subtitles of movie series, and (c) the ParaCrawl corpus (Esplà et al., 2019) consisting of automatically crawled translations from document-level translations of European Parliament Proceedings.

As expected, our model performs minimal editing on the high-quality *Tatoeba* bitexts. Concretely, only $\sim$ 30% of it gets revised, while as suggested by the token-level TER statistics, even the revised sentence pairs mostly consist of substituted tokens. Further manual inspection reveals that most of those tokens depict subtle spelling differences between Greek words. On the other hand, when editing the samples of automatically extracted bitexts, our refinement model performs more frequent edits: it revises $\sim$ 65% of OpenSubtitles and $\sim$ 88% of ParaCrawl bitexts. Interestingly, although a greater amount of ParaCrawl texts get revised compared to OpenSubtitles, edits on the latter are more aggressive as it consists of at least 10% fewer "correct" (i.e., **C**) tokens than the former. A break of the types of operations further reveals that editing OpenSubtitles requires more "deletion" (i.e., **D**) and "insertion" (i.e., **I**) operations compared to the other two. This observation connects to prior efforts on auditing OpenSubtitles that found sentence segmentation errors (i.e., added extra leading/trailing words on one side) to be a frequent type error for this corpus (Vyas et al., 2018a).

5.3.5 Summary

We introduce an alternative approach for bitext quality improvement that we show is better suited for low-resource language pairs. Instead of filtering out imperfect translation

references that result from automatic bitext mining, we instead edit them with the goal of improving their quality. Our editing models are trained using only synthetic supervision, which can be gathered at scale for any language pair that support bitext mining. Extensive quantitative analysis suggests that our approach successfully improves bitext quality for a variety of language pairs and different resource conditions. Furthermore, extrinsic experiments on 10 low-resource NMT tasks suggest that bitext refinement constitutes a successful approach to improving NMT translation quality in low data regimes. Those findings highlight the importance of the good *quality* bitexts in scenarios where large *quantities* cannot be guaranteed and motivate future research on improving low-resource NMT further.

5.4 Conclusion

This section closes the loop of our exploration of how fine-grained divergences interact with MT training. We explored two orthogonal directions to mitigate the negative impact of semantic divergences on NMT systems. The first direction relies on detected divergences in training bitext and aims to model them explicitly via challenging the assumptions of translation equivalence adopted by the standard NMT framework. In doing so, we introduced DIV-FACTORIZED, a divergent-aware NMT framework that models divergences as factors via encoding information about which tokens are indicative of semantic divergences between the source and target side of a training sample. Our empirical results indicate that factors to help NMT recover from the degradation caused by naturally occurring divergences, improving both translation quality and model calibration on MT tasks.

Our second direction aims, instead, to equivalize divergences in bitext in order to match the NMT translation equivalence assumptions better. In doing so, we first contribute an empirical study that explores how synthetic translations can equivalize semantic divergences in mined bitexts under semantic control signals. We show that selective replacement of

divergences with synthetic translation candidates is a successful strategy to bitext quality improvement in settings where enough data is available to build NMT systems of reasonable quality. In settings where the latter cannot be guaranteed, we propose BITEXTEDIT, an edit-based model that refines divergences in mined bitexts in an autoregressive fashion via utilizing signals from both source and target references more effectively than selective replacement with MT.

After establishing that the detection of fined-grained divergences can help us improve machine translation systems, we now ask: ***How can the detection of semantic divergences advance humans' understanding of translation differences?*** Our last and final chapter aims to answer this question by proposing an approach that explains divergence predictions and evaluating the usefulness of such predictions in real-world use cases.

Chapter 6: Assisting Humans to Detect Translation Differences by Contrastive Phrasal Highlighting

In our final piece of work, we hypothesize that in order for translation to break language barriers, we need NLP tools that not only compare and contrast the meaning of texts across languages but also explain their differences to their consumers. As a result, our primary goal in this chapter is to equip our own divergence detector models with the ability to indicate not just *when* divergences exist but also tell us *where* such divergences reside.

EL Τα μέτρα που εφαρμόστηκαν στην Ελλάδα δέχθηκαν πολλά θετικά σχόλια από τον ξένο τύπο που εξήρε την Ελλάδα για τον περιορισμό της εξάπλωσης της πανδημίας στη χώρα.
GLOSS The measures put place in Greece received many positive comments from the foreign press for having slowed the spread of the disease in the country.
EN The measures put in place in Greece are among the most proactive and strictest in Europe and have been credited internationally for having slowed the spread of the disease.

Figure 6.1: Highlighted meaning differences in Greek (EL) and English (EN) sentences drawn from Wikipedia translated articles. Highlights show phrases that differ in meaning across the two languages, and color indicates which Greek phrases should be compared with which English phrases.

To that end, we devise an explainability approach drawing from a key insight from social sciences: Human explanations are contrastive (Miller, 2019), i.e., humans barely explain why an event happened in the vacuum, but rather why it happened compared to a contrast case. For instance, to explain how a Greek sentence and a candidate translation in English differ, it is more informative to show how a phrase pair in one language contrasts with a phrase in the other language (as illustrated in Figure 6.1) than to highlight all salient Greek and English tokens without specifying how they relate to each other.

Unlike prior work in explainability that seeks to explain what tokens led to a model prediction, we ask: *"What differences between the two inputs explain a prediction?"* In what follows, we start by reviewing prior work on explainability (§6.1), then introduce our approach (§6.2) and finally turn into evaluating our approach based on proxy (automatic) (6.3) and application-grounded evaluations (§6.4 and §6.5).

6.1 Background

Explainability approaches aim at explaining models' predictions at a global level or a local level (Doshi-Velez and Kim, 2017). The former approaches analyze the general patterns that help us understand a model's global behavior, while the latter explain why a prediction related to a specific input was made. Our work draws on work on local explanations, as detailed below.

6.1.1 Highligh-based Explanations

In general, a popular approach to local explanations aims to highlight features in a model's input that are deemed important for a specific prediction. Such highlight-based explanations can take different forms, for instance, attributing feature importance based on saliency maps (Smilkov et al., 2017; Shrikumar et al., 2017; Sundararajan et al., 2017; Adebayo et al., 2018). Drawing on such approaches, a significant amount of NLP work joined the debate on whether deep neural models' internal mechanisms, such as gradients or attention weights, can offer modules to communicate a notion of input tokens' importance implicitly (Bastings and Filippova, 2020; Bibal et al., 2022). While the debate is still ongoing, some works contribute evidence questioning their applicability to serve as an explanation framework (Jain and Wallace, 2019; Wang et al., 2020a) and others suggesting technical solutions and evaluations under which attention *could* be seen as an explanation (Wiegreffe

and Pinter, 2019; Tutek and Snajder, 2020; Sun and Marasović, 2021). On the other hand, another line of NLP research adapts popular erasure-based techniques to explaining model predictions by treating them as black boxes. For instance, Ribeiro et al. (2016) approximates the model's local decision boundary with linear models by erasing multiple tokens from a mode's input at random. Li et al. (2016) compute a token's importance by measuring the difference in the model's confidence once it is erased from the input, with several following works exploring different token erasure schemes, based on input reduction (Feng et al., 2018) or input marginalizing (Kim et al., 2020). Drawing on this line of work, instead of attributing token importance based on erasure operations, we erase instead *phrasal pairs* that together contribute to the model's prediction interdependently.

6.1.2 Contrastive Explanations

Contrastive explanations have recently received attention within the NLP literature in light of insights from social science research emphasizing that human explanations are usually perceived with respect to an implicit contrast case (Miller, 2019). To that end, several approaches have been introduced aiming at *generating* such implicit contrast cases, namely counterfactuals. Within the currently studied NLP tasks, counterfactual explanations are usually conceptualized as minimal textual edits on an input instance that cause a flip on the model's prediction to a contrast class. Performing such edits usually requires training dedicated editor models (Ross et al., 2021), prompting language models (Paranjape et al., 2021), and accessing external knowledge bases (Chen et al., 2021), among others (Li et al., 2020; Chemmengath et al., 2022). Unlike the above approaches that *generate* contrastive explanations, a few studies propose to extract *highlights* contrastively. Concretely, Jacovi and Goldberg (2021) proposed to extend erasure-based approaches based on the inclusion of contrastive features, i.e., identify what parts of the input lead to the prediction of the main

class and what parts lead to the prediction of a contrast class, or by extending saliency-based methods to explain the generation of contrast cases for language generation tasks Yin and Neubig (2022). Contrary to prior work, we study how contrastive explanations can be operationalized to surface salient phrasal pairs contributing to NLP prediction of regression models.

6.1.3 Evaluation of Explanations

Evaluating models' explanations is an active research area. Most current work in NLP adopts *proxy* evaluation approaches comparing the automatically extracted explanations with human-provided rationales, i.e., tokens within a text deemed important for the gold label. To that end, numerous datasets have been released (Wiegreffe and Marasovic, 2021) covering a wide range of explanations types, such as highlgihts (Zaidan et al., 2007a; Khashabi et al., 2018; Thorne et al., 2018; DeYoung et al., 2019) or explanations in natural language (Camburu et al., 2018). Other approaches rely on *simplified* tasks such as simulatability experiments, where users predict the model's output by accessing only its explanation (Hase and Bansal, 2020; Nguyen, 2018). Despite the attractiveness of proxy and simplified evaluations, work on the intersection of human-computer interactions, and AI raises concerns about their reliability and warns that conclusions on the effectiveness of explanations based on such evaluations might be misleading (Buccinca et al., 2020). Along those lines, Boyd-Graber et al. (2022) suggest that evaluation of NLP explanations should be based on application-grounded evaluations that aim to understand the use of explanations by humans in completing a concrete task, as suggested by work in HCI (Doshi-Velez and Kim, 2017; Suresh et al., 2021; Liao et al., 2020). Our work draws insights from HCI and, unlike prior evaluation on NLP explanations, complements complementing proxy evaluations with application-grounded user studies.

6.1.4 Detecting & Explaining Translation Differences

Detecting translation differences has been long studied in the context of analyzing and understanding human and machine translations. Broadly speaking, most approaches fall under the broad categories of assigning sentence-level scores indicative of translation quality (Zhang et al., 2019; Sellam et al., 2020; Rei et al., 2020; Xu et al., 2022) or detecting meaning divergences (Vyas et al., 2018b; Briakou and Carpuat, 2020; Zhai et al., 2019a; Wein and Schneider, 2021). More recently, several task framings that look into the detection of translation differences at a finer granularity have been proposed, such as phrasal detection of translation processes (Zhai et al., 2018) and word-level quality estimation (QE) tasks that comprise standardized shared tasks in WMT since 2018 (Specia et al., 2018; Fonseca et al., 2019; Specia et al., 2020, 2021). Amongst the latter, explainability approaches have been proposed to predict word-level errors in machine translation outputs based on Explainable QE shared tasks (Fomicheva et al., 2021; Zerva et al., 2022), Despite the success of leading approaches (Treviso et al., 2021; Rei et al., 2022) based on proxy evaluations, they are primarily based on saliency-based methods that do not adhere to the contrastive nature of human explanations, while little is known as to whether such they are easy to comprehend and helpful to humans. Our work evaluates our proposed contrastive phrasal highlights with human-centered evaluations grounded on real-world scenarios.

6.2 Explaining Divergences with Contrastive Phrasal Highlights

We introduce a highlighting approach to explain the predictions of a model that compares and contrasts two text inputs. Since models operating on top of two textual inputs require capturing the relationships of content residing across them, we propose to explain such models by surfacing their cross-sentential relationships as contrastive phrasal highlights. We

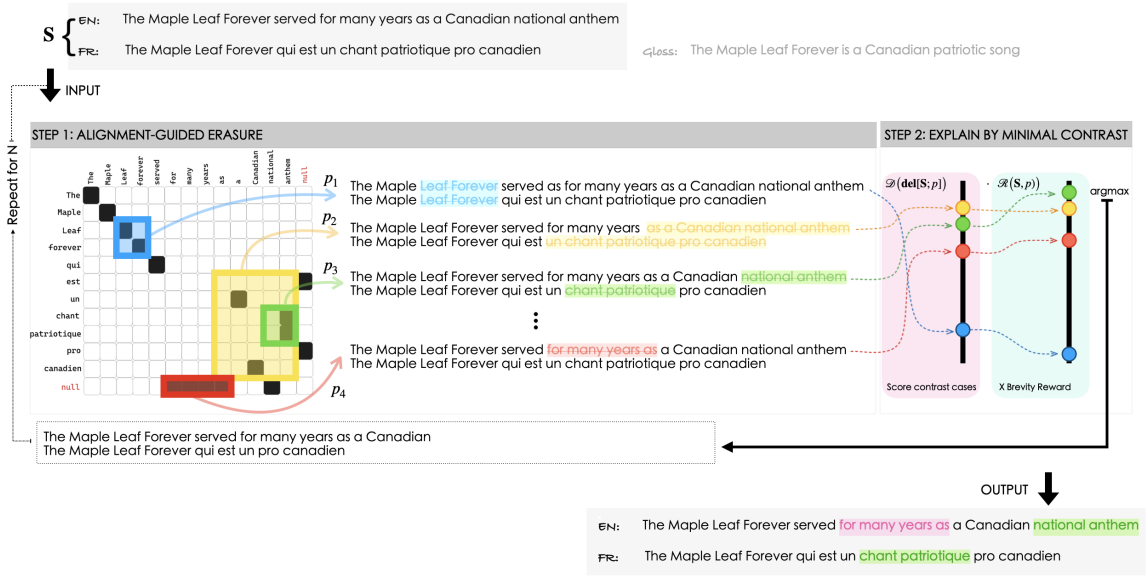


Figure 6.2: Explain by contrastive phrasal highlights: Our approach takes as input a sentence pair ($\mathbf{S}$) and extracts a set of perturbed inputs by erasing phrasal pairs guided by word alignments (Step 1). Then it explains the prediction of a regressor $\mathcal{D}(\mathbf{S})$ by highlighting the phrasal pair p that, once deleted, maximizes the model's prediction $\mathcal{D}(\mathrm{del}[\mathbf{S};p])$ multiplied by a brevity reward $\mathcal{R}(\mathbf{S},p)$ that encourages the extraction of short phrasal pairs (Step 2).

hypothesize that phrasal-(pairs) is a more appropriate definition of *cognitive chunks* (Doshi-Velez and Kim, 2017), i.e., the basic units of explanations, compared to tokens. For instance, suppose we want to explain why the English and French texts in Figure 6.2 differ in meaning. The English text describes the Maple Leaf Forever as a *national anthem* while the French text as a *chant patriotique* (patriotic song). As a result, those two segments exhibit an interdependent relationship in explaining the differences between the two texts and should be highlighted together. On the other hand, the added phrase *"for many years"* in English is independently highlighted as when it contrasts with the French text, no direct correspondence can be established.

Our approach is motivated by erasure-based explanation methods (Ribeiro et al., 2016; Lundberg and Lee, 2017), which identify salient parts of an input based on how much erasing them impacts the prediction of the NLP model. However, instead of independently erasing

tokens from the two input texts, we generate counterfactual inputs by erasing contrastive phrase pairs. Here, we apply this approach to explaining the predictions of our own divergent model introduced in Chapter 3. As a reminder, our divergent model ranks bilingual sentence pairs based on the granularity of their semantic divergences (assuming x is an equivalent pair and $\hat{x}$ a pair containing semantic divergences, our model $\mathcal{R}(.)$ ranks them such that $\mathcal{R}(x) > \mathcal{R}(\hat{x})$. We introduce a lightweight, iterative two-stage approach as shown in Figure 5.5. Given two input texts, we first extract a set of candidate counterfactual inputs based on an alignment-guided erasure approach that masks out aligned segments (§6.2.1). Then, we seek to explain the model's prediction by selecting the phrase pair whose erasure maximally increases the similarity score between the two inputs (§6.2.2).

6.2.1 Alignment-Guided Phrasal Erasure

We propose an alignment-guided erasure approach that takes into account the input's structure. Given a sentence-pair $\mathbf{S}$, we start by extracting a set of candidate counterfactual instances by deleting a single phrase(-pair) from $\mathbf{S}$. Given that erasing all possible phrase(-pairs) for each sentence is computationally prohibited, we restrict our search to deleting phrases that belong to an aligned phrase table, $\mathcal{P}$.

Our phrase pair candidates for erasure are derived from classical statistical machine translation techniques developed for extracting phrase translation dictionaries from parallel corpora (Och et al., 1999; Koehn et al., 2007). Given two texts, we derive word alignments based on multilingual word embeddings (Jalili Sabet et al., 2020) and then extract all phrase pairs consistent with the word alignments. A phrase pair (p_1, p_2) is consistent with the word alignment if p_1 and p_2 are each contiguous sequences of words in each language if there exists at least one alignment link that connects a word within p_1 to a word within p_2, and where all alignment links that start within p_1 map to tokens within p_2 and vice versa.

Unaligned words (i.e., words aligned to a `null` token) can be included in either phrase. As a result, the extracted phrase pairs comprise not only equivalent phrases that are exact translations of each other but also include related but divergent phrase pairs, such as such as the aligned phrases *"national anthem"* versus *"chant patriotique"*, and the unaligned phrase *"served for many years"* (see Figure 6.2).

6.2.2 Explain by Minimal Contrast

In the second stage, given the sentence pair $\mathbf{S}$ and its aligned phrase table $\mathcal{P}$, we first extract a set of contrast cases $\tilde{\mathcal{P}} = \{p\}$ consisting of all the phrasal pairs that, once erased, make the two inputs more equivalent, as measured by an increase in score larger than a margin ϵ:

$$\tilde{\mathcal{P}} = \Big\{ p \in \mathcal{P}, \;\; \text{s.t.} \;\; \mathcal{R}\Big(\textbf{DEL}\big[\mathbf{S}; p\big]\Big) > \mathcal{R}\Big(\mathbf{S}\Big) + \epsilon \Big\}$$

Then, we extract a contrastive phrasal highlight that explains the prediction of the model, by maximizing the below product:

$$\text{argmax} \left\{ \mathcal{R}\Big(\textbf{DEL}\big[\mathbf{S}; p\big]\Big) \cdot \mathcal{BR}\Big(\mathbf{S}, p\Big) \right\}$$

$$\mathcal{BR}(\mathbf{S}, p) = \begin{cases} e^{-\frac{|p|}{|\mathbf{S}|}}, & \text{if } \mathcal{R}\Big(\textbf{DEL}\big[\mathbf{S}; p\big]\Big) \geq 0 \\ e^{+\frac{|p|}{|\mathbf{S}|}}, & \text{otherwise} \end{cases}$$

where the first term $\mathcal{R}(\textbf{DEL}\big[\mathbf{S}; p\big])$ encourages the extraction of a contrastive phrasal highlight p corresponding to a high score under the model (i.e., deleting this phrase pair yields a contrast case), while the second term, $\mathcal{BR}\big(\mathbf{S}, p\big)$ corresponds to a *brevity reward* that encourages extraction of shorter highlights. The latter component operationalizes the idea

that explanations should be minimal (Hilton, 2017; Lipton, 1990), covering only the most relevant causes to reduce cognitive load. Therefore, we seek to explain the prediction through minimal deletion edits.

The above approach yields one phrasal(-pair) that explains the divergence prediction for the original sentence $\mathbf{S}$. To extract multiple potential explanations, we repeat this process iteratively by erasing the extracted contrastive phrasal highlight from the current input sentence $\mathbf{S}' = \textbf{DEL}\big[\mathbf{S}; p\big]$, and repeat the steps outlined in §6.2.1 and §6.2.2. This iterative process ends when, for a given input, none of the extracted counterfactuals instances yield a more equivalent pair, i.e., $\tilde{\mathcal{P}} = \emptyset$, or we reach an equivalent input under the divergent ranker, i.e., $\mathcal{R}[\mathbf{S}'; p] > 0$.

6.3 Proxy Evaluation

In this section, we describe our proxy evaluation based on human-provided highlights. We acknowledge that proxy evaluations encounter validity issues (Boyd-Graber et al., 2022) and use them primarily to guide system development and validate against standard highlighting methods.

6.3.1 Experimental Setup

Explanandum We seek to explain the prediction of our divergence ranking model that assigns a continuous score to a sentence pair, reflecting its divergence. For more information on the model details and the experimental setup, we refer the reader to §3.2.

Explainers We contrast our contrastive phrasal highlights against two standard post-hoc explanation methods that seek to explain the predictions of the explanandum detailed above.

Concretely, we use LIME (Ribeiro et al., 2016), which fits a linear model in the proximity of each instance, to approximate the behavior of the explanandum, and SHAP (Lundberg and Lee, 2017), which computes Shapley values by estimating the marginal contribution of each word across all possible perturbations.

HUMAN The older generation turbines generate kilowatts , and the modern turbines installed generate up to 3 megawatts , depending on the specific turbine and manufacturer .. Les turbines d' ancienne génération génèrent des kilowatts , alors que les éoliennes modernes ont une puissance pouvant aller jusqu' à 3 mégawatts . *(gloss) The turbines of the old generation generate kilowatts, while the modern wind turbines generate up to 3 megawatts of power.*
LIME The older generation turbines generate kilowatts , and the modern turbines installed generate up to 3 megawatts , depending on the specific turbine and manufacturer .. Les turbines d' ancienne génération génèrent des kilowatts , alors que les éoliennes modernes ont une puissance pouvant aller jusqu' à 3 mégawatts .
SHAP The older generation turbines generate kilowatts , and the modern turbines installed generate up to 3 megawatts , depending on the specific turbine and manufacturer .. Les turbines d' ancienne génération génèrent des kilowatts , alors que les éoliennes modernes ont une puissance pouvant aller jusqu' à 3 mégawatts .
OURS The older generation turbines generate kilowatts , and the modern turbines installed generate up to 3 megawatts, **depending on the specific turbine and manufacturer ..** Les turbines d' ancienne génération génèrent des kilowatts , alors que les éoliennes modernes ont une puissance pouvant aller jusqu' à 3 mégawatts .

Table 6.1: Examples of divergence explanations (**HUMAN** corresponds to REFRESD rationales).

Human-Provided Highlights We evaluate our approach on our Rationalized English-French Semantic Divergences (REFRESD) dataset, which is annotated with divergences of fine-grained and coarse-grained granularity, along with rationales that justify the sentence label. We compare our contrastive phrasal highlights with the human-provided highlights of instances annotated as having "Some Meaning Differences" at the sentence level, where we expect the more subtle differences in meaning to be found and the phrasal annotation to be most challenging. As a reminder, the evaluation set for this class comprises 418 English-French sentence pairs. For more details on the dataset, we refer the reader to §3.1.

Evaluation Metrics Motivated by prior work on contrastive explanation generation (Ross et al., 2021) and evaluation of rationalized NLP models (DeYoung et al., 2020b), we

	ENGLISH			FRENCH		
	PR.	RE.	DEL.	PR.	RE.	DEL.
ORACLE	100	100	39%	100	100	43%
Random	39	48	50%	42	49	50%
LIME	45	37	30%	44	37	34%
SHAP	53	34	25%	50	32	26%
OURS (− BR)	52	76	54%	56	74	55%
OURS	58	61	37%	60	55	37%

Table 6.2: Proxy Evaluation Results with respect to gold-standard rationales on the “Some Meaning Difference" classes of REFRESD.

compute two measures: a) **Agreement with Human Rationales**: the Precision, Recall scores averaged across instances for each side of the sentence-pair computed against the gold-standard rationales in REFRESD; b) **Minimality**: the length of the contrastive phrasal highlights measured as the number of tokens.

6.3.2 Results

Table 6.3 presents the results of our proxy evaluation.

Comparison with Baselines Explaining the explanandum’s prediction by extracting contrastive phrasal highlights significantly outperforms both LIME and SHAP based on Precision, Recall scores on REFRESD. Interestingly, the standard highlighting baselines fail to perform better than the RANDOM baseline.[1] A closer look at the highlights produced by the different approaches, as shown in Table 6.1, indicates that both LIME and SHAP suffer from two major issues: sparsity and accuracy. We hypothesize that those issues are linked to the fact that both approaches operate in the token space, by employing random token masking and ignoring the interdependent relationships between segments residing across languages. By explicitly modeling such relationships, our approach produces highlights that

[1]Note that LIME and SHAP are feature attribution methods that assign a continuous score to each word. Changing the threshold to different values than the default $t = 0$ does not improve the results, as both methods suffer from a precision-recall tradeoff.

match the ones in REFRESD better.

Ablating the Brevity Reward Across variants (i.e., with and without the brevity reward), we extract explanations that match rationales in REFRESD better than all baselines. The unconstrained variant of our approach (i.e., Ours (-BR)) achieves the highest recall. However, it does so by hurting precision (i.e., it produces fewer and longer highlights per instance). Concretely, on average, it highlights more than 50% of each sentence-pair, while the oracle average length base on gold standard highlights corresponds to about 40%. Using the brevity reward results in highlights that, on average, match the length of the gold standard highlights better, indicating that this factor indeed helps produce more minimal explanations.

In summary, the proxy evaluation results and manual inspection are promising indicators that contrastive phrasal highlights could be helpful when detecting semantic divergences. However, as highlighted by recent work on explanation evaluation, proxy evaluations can be misleading. In what follows, we present results on two human-centered evaluations of our proposed approach, where we aim to evaluate whether explanations are helpful to humans in two application-grounded scenarios. Note that both of the following studies have been approved by the University of Maryland Institutional Review Board (IRB number 2018458-1).

6.4 Application-Grounded Evaluation I: Annotation of Semantic Divergences

Given that the premise of our work is to produce explanations that serve human needs, we conduct a human-centered evaluation and explore whether contrastive phrasal highlights assist bilingual speakers in detecting fine-grained meaning differences in bilingual texts. As

shown by prior work, including ours, annotating fine-grained divergences is a challenging task and requires dedicated annotation protocols based on rationales (§3.1) or abstract meaning representation frameworks (Wein et al., 2022) to achieve moderate agreement. As a result, such annotations are usually hard to collect with crowd workers since they require explicit annotator training. In this section, we attempt to understand whether providing bilingual crowd workers with contrastive phrasal highlights can help collect reliable annotations of divergence and potentially ease the need for complex and expensive annotation protocols. To that end, our goal is to test the following hypothesis: *Contrastive phrasal highlights improve the annotation of fine-grained semantic divergences in terms of accuracy and agreement*. In what follows, we describe the experimental setup of our study (§6.4.1), then detail our study design (§6.4.2), and conclude with findings (§6.4.4).

6.4.1 Experimental Setup

Explanandum We seek to explain the predictions of divergentmBERT. We train separate models for English-French and English-Spanish following the same approach to acquiring synthetic supervision based on WikiMatrix data, as detailed in §6.3.1.

Study Data We construct a synthetic dataset of fine-grained divergences that mimic translation processes used by human translators (Zhai et al., 2019a). To create this dataset, we use translations in English-French and English-Spanish drawn from the FLORES evaluation set (Goyal et al., 2022), which consists of human-provided references of English Wikipedia sentences. We note that FLORES is a multi-parallel resource and therefore enables a controlled comparison between the two languages studied by focusing on the translations of the same set of English sentences. Moreover, the motivation behind introducing a synthetic dataset is to ensure that divergences resemble the ones arising during the translation processes and factor out other potential coarse divergence phenomena that might appear in

currently annotated divergences that are curated from parallel corpora (Vyas et al., 2018b; Kreutzer et al., 2022a). To that end, we first identify semantically equivalent pairs using divergentmBERT in English-French and English-Spanish translations. In what follows, we call those translation pairs *seeds*. Then starting from a random sample of 50 seeds that share the same English reference across the two languages, we introduce fine meaning differences by editing the English references of 10 of those samples. We perform edits that resemble translation processes, such as modulation, explicitation, and reduction (Zhai et al., 2019a), by online interacting with chatGPT[2]. Through online interaction, one of the authors reviews the edits generated by the model to ensure that we only introduce fine-grained divergences that could reflect human translation processes. For instance, we introduce a generalization process as follows:

> Despite leaving the show in 1993 he kept *[the title of executive producer]* → *[a senior role]* and continued to receive tens of millions of dollars every season in royalties.

We include more examples of the introduced semantic divergences in Figures 6.5 and 6.6.

Finally, to ensure that the detection of semantically equivalent pairs is not a trivial task for humans, we additionally introduce syntactically divergent translations. To that end, we paraphrase the English reference of 10 seeds by again interactively reviewing and revising chatGPT edits to ensure the introduced edits are constrained to syntactic phenomena and do not introduce changes in meaning. We find that such paraphrase-based perturbations usually "confuse" the explanandum, leading to the extraction of false positive highlights. We include examples of the introduced syntactic divergences in 6.7. In total, the synthetic dataset consists of 35 semantic equivalent translations (split into 25 original translations and 10 syntactically divergent paraphrases) and 15 fine-grained divergences resembling translation processes.

[2] https://chat.openai.com/

6.4.2 Study Design

To test our hypothesis, we ran a controlled evaluation across two language pairs, i.e., English-French and English-Spanish. We describe the task and study design details below.

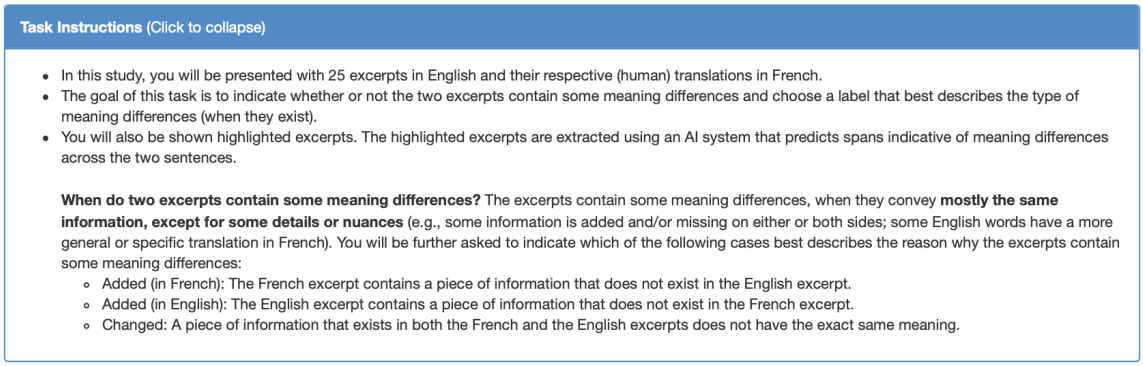


Figure 6.3: Instructions for application-grounded evaluation I: Annotation of Semantic Divergences.

Task Description An annotation instance is a sentence-pair in English and, e.g., French. Bilingual speakers are asked to read each sentence pair closely and determine whether the *"sentence pair contains any difference in meaning"*. Furthermore, they are asked to characterize the difference as conveying added information ("Added") or information that is present in both languages but does not precisely match ("Changed") to mirror prior annotation divergence protocols (Briakou and Carpuat, 2020). We include a screenshot of the task description presented to annotators in Figure 6.3.

Conditions We study two conditions. In the first, participants are shown highlights, while in the other, they are not (✓ HIGHLIGHTS vs. ✗ HIGHLIGHTS). The only information available to participants about the highlights is they *"are AI-generated and indicative of meaning differences"* as we want them to choose how and whether they use them in their assessments based on their own intuitions. We corroborate that highlights are not presented

as explanations for divergence predictions as we do not provide them with the latter.

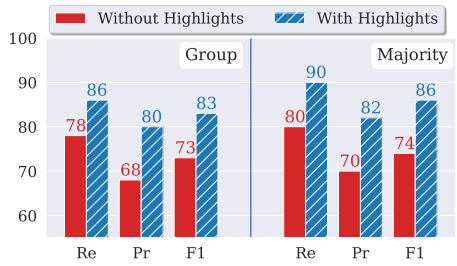
Procedures We conduct a between-subjects study where participants are randomly assigned to a condition. Participants are first presented with a tutorial that explains the task and relevant terminology. Each participant is presented with 25 instances from one of the two studied conditions. Each batch of annotated instances contained 30% of semantically divergent edits, 20% of syntactically divergent edits, and 50% original semantically equivalent pairs. Instances within each batch are randomized. We include two attention checks where participants are asked to indicate their answers to the previous question. After completing the task, participants were asked to complete a brief survey assessing their perception of the AI-generated highlights (if assigned to this condition) and finally were asked to provide their free-form feedback on the study along with demographic information, such as gender and age. The average time of the study was 20 minutes. In sum, we collect 3 annotations per instance and annotate a total of 100 instances (50 per condition) for each of the two language pairs studied. We include screenshots of the annotation task interface in Figure 6.6.

Participants We recruit 12 participants per language pair using Prolific.[3] Each participant is restricted to taking the study only once. None of the participants failed both attention checks; hence we did not exclude any of the collected annotations from our analysis. All participants self-identified as bilingual speakers in the two languages involved in each study. Participants are compensated at a rate of 15 USD per hour.

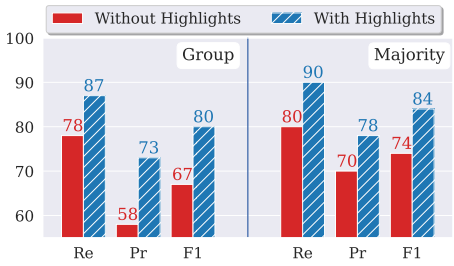
6.4.3 Measures

Our main evaluation measures are Precision, Recall, and F1 computed by assuming that semantically edited instances correspond to divergences, while the rest are treated as

[3] https://www.prolific.co/



(a) English-French



(b) English-Spanish

Figure 6.4: Annotation-grounded evaluation results without and vs. with highlights. Contrastive highlights improve the annotation of fine-grained semantic divergences. All improvements are statistically significant ($p = 0.1$).

semantically equivalent pairs. We report those accuracy statistics both at the group level and also by majority voting annotations. Furthermore, we summarize the responses provided as free-form feedback and report subjective measures that reflect participants' perceived understanding of the explanations, i.e., *"the highlights are useful in detecting meaning differences"* if provided. The latter measures are collected on a 5-point Likert scale.

6.4.4 Study Results

Reliability of Annotations As shown in Figure 6.4, the bilingual *group* that annotated fine-grained meaning differences in the presence of contrastive explanations—✓ HIGHLIGHTS, is significantly ($p = 0.1$) more accurate, across Precision, Recall, and F1 scores and language-pairs, compared to the group that does not access them—✗ HIGHLIGHTS. Furthermore, those improvements carry over when aggregating annotation results by *majority* voting annotations across instances. As a result, this leads us to accept the hypothesis that contrastive phrasal highlights improve the annotation of fine-grained divergences by bilingual speakers. Finally, detecting divergences with highlights additionally improves the reliability of annotations are measured by Cohen's cappa inter-annotator agreement statistics:

	EN-FR	EN-ES
✗ HIGHLIGHTS	51 (moderate)	33 (fair)
✓ HIGHLIGHTS	66 (substantial)	52 (moderate)

Subjective Measures & User Feedback Overall, bilingual speakers presented with contrastive phrasal highlights agreed they were useful in helping them spot fine-grained divergences—average self-reported usefulness of 3.8 for EN-FR and 4.2 for EN-ES. Finally, although contrastive phrasal highlights were useful, annotators also note that they cannot entirely rely on them. For instance, some of the participant's feedback is *"The highlights are useful but not* 100% *reliable, that is because I found other added/changed words that AI did not highlight"* and *"In most cases, the words highlighted by the AI have helped to detect possible differences"*.

(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 6.5: Annotations of fine-grained semantic divergences (English-Spanish on the left and English-French on the right) reflective of *explicitation* and *reduction* translation processes. Contrastive highlights subsume added content presented in one language but missing from the other.

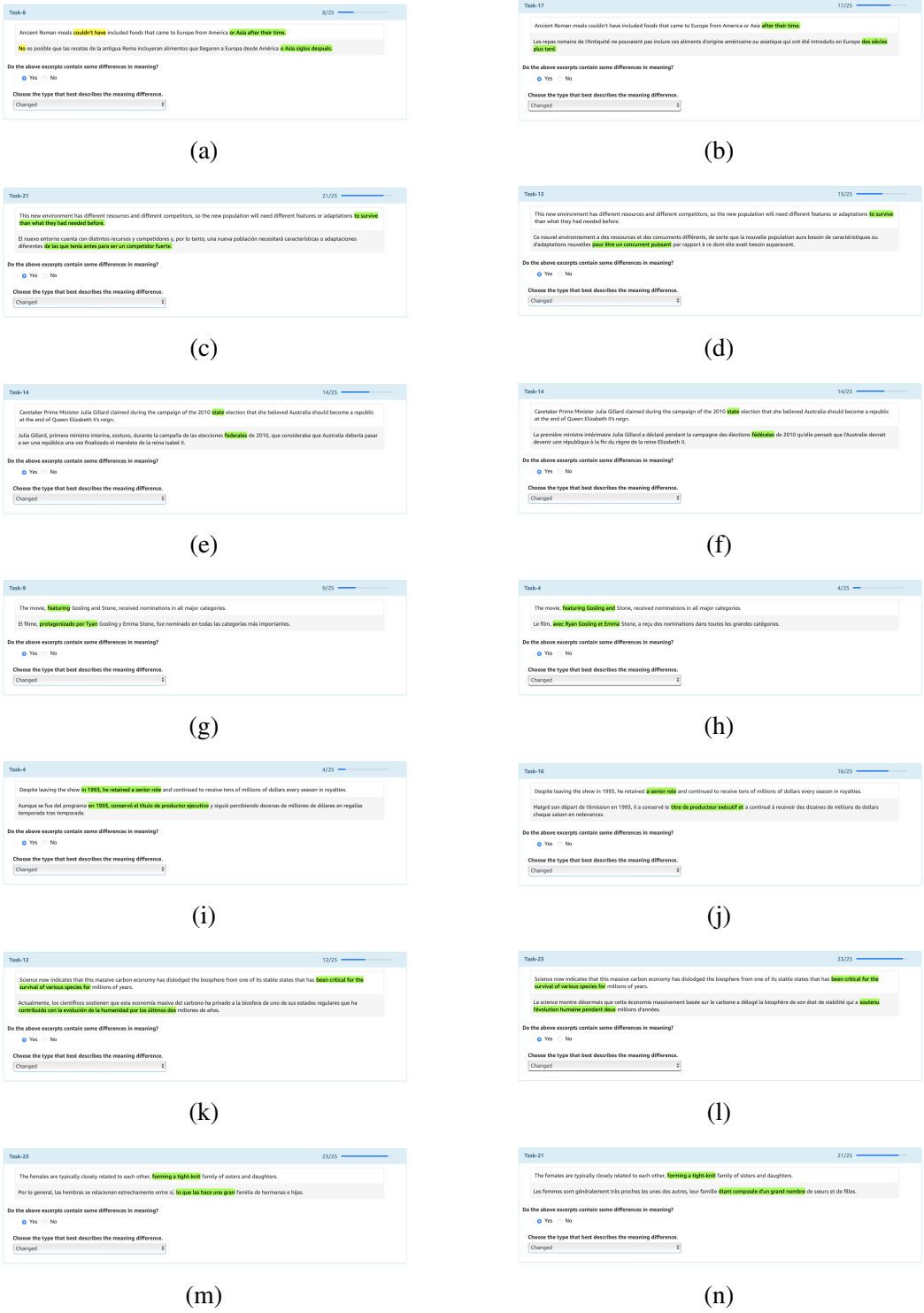


Figure 6.6: Annotations of fine-grained semantic divergences (English-Spanish on the left and English-French on the right) reflective of *modulation* translation processes. Contrastive highlights frequently subsume content that the is does not contain the exact same meaning across languages.

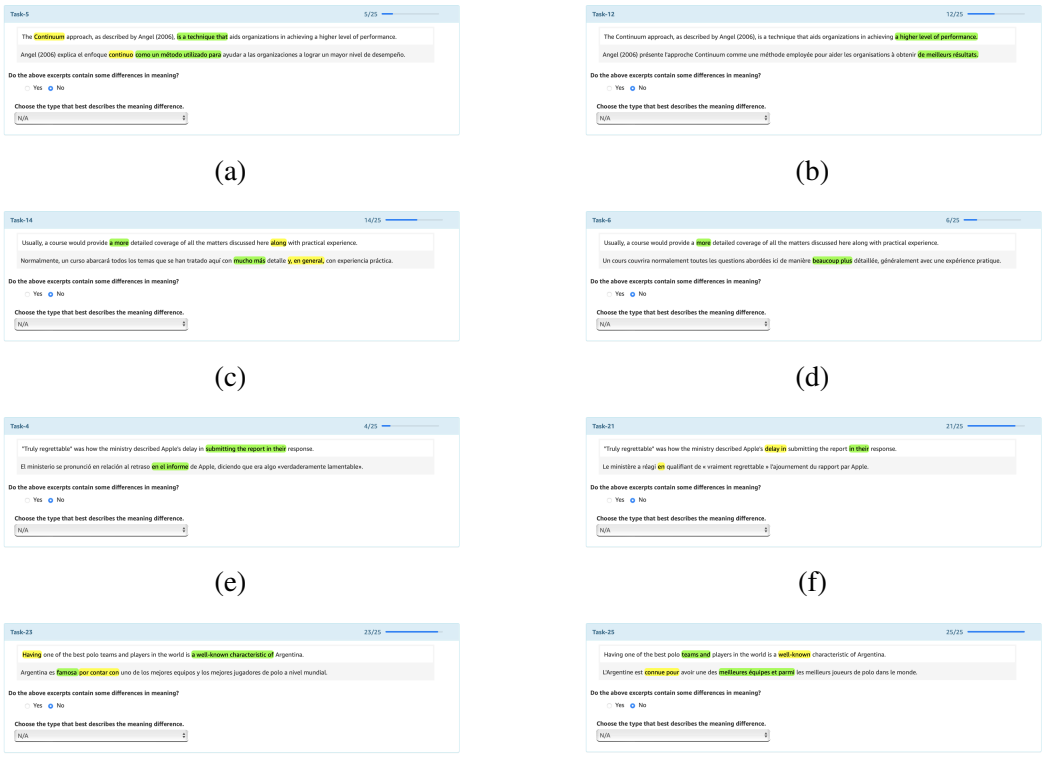


Figure 6.7: Annotations of semantically equivalent pairs (English-Spanish on the left and English-French on the right) that reflect syntactic divergences. Contrastive highlights surface false positive segments.

6.5 Application-Grounded Evaluation II: Critical Error Detection

In this section, we examine the potential of using contrastive phrasal highlights to assist bilingual humans in detecting critical errors in machine translation outputs. Recent work on quality estimation for machine translation proposes an error-based evaluation framework where bilingual *professional* annotators are asked to: highlight translation errors; rate their severity (i.e., major vs. minor); and finally, indicate their type (e.g., mistranslation error) (Freitag et al., 2021b). Drawing on those intuitions, we aim to study whether a simplified evaluation framework that uses contrastive phrasal highlights yields reliable annotations of critical errors with bilingual *crowd workers*. This is a hard task as these errors are rare in high-quality systems and might be missed when reading too fast. In what follows, we test the hypothesis: *Contrastive phrasal highlights help bilingual crowd-workers detect critical errors in machine translation outputs*. We start by describing the experimental setup (§6.5.1), then turn to detail the study design (S6.5.2), and conclude with findings (§6.5.3).

6.5.1 Experimental Setup

Explanandum We seek to explain the divergent predictions of divergentmBERT trained for English-Portuguese following the process described in §6.3.1. We emphasize that the explanandum is sensitive to detecting any meaning differences and not only critical errors, and therefore we expect the contrastive phrasal highlights to surface both minor and major accuracy errors.

Study Data For the purposes of the study, we use the synthetic Portuguese-English dataset from the Critical Error Detection task of WMT (Freitag et al., 2021c). The dataset consists of pairs sampled from the OPUS parallel corpus (Tiedemann, 2012a) treated as "not containing a critical error". Then, the WMT organizers artificially corrupted 10% of

those pairs to introduce critical errors that reflect hallucinated content, untranslated content, or mistranslated texts. We will refer to those errors as WMT for the rest of the paper. Additionally, to make the task hard for humans, we a) filter out WMT critical errors that concern deviation in numbers, time, units, or dates and b) we artificially introduce *negation* errors: we edit the English text locally by changing one or two words in such a way that the meaning of the sentence is entirely flipped (e.g., the word *benefits* in an English reference is replaced by the word *harms*). In total, the study dataset consists of 30 translations that do not contain a critical error (also detected as equivalent under the explanandum), 10 translations reflecting WMT errors, and 10 translations reflecting local *negation* errors.

6.5.2 Study Design

We test our hypothesis with a user study of critical error detection in English translations of Portuguese texts, as described below.

Task Description An annotation instance is an excerpt in Portuguese and its (machine) translation in English. First, bilingual speakers are asked to read the two excerpts and determine whether the translation contains an accuracy error. Concretely, we ask *"Does the translated excerpt contain any error in meaning?"* If an accuracy error is detected, we ask them to rate its severity as being *minor* or *major*. Minor errors are defined as errors that do not lead to loss of meaning and would not confuse or mislead the reader, while major ones may confuse or mislead the reader due to significant changes in meaning. We ask participants to factor out fluency issues in their assessment as much as possible. We include a screenshot of the task description presented to annotators in Figure 6.8.

Conditions & Evaluation Constructs We use Precision, Recall, and F1 as our main evaluation measures and follow the same condition treatments and subjective measures

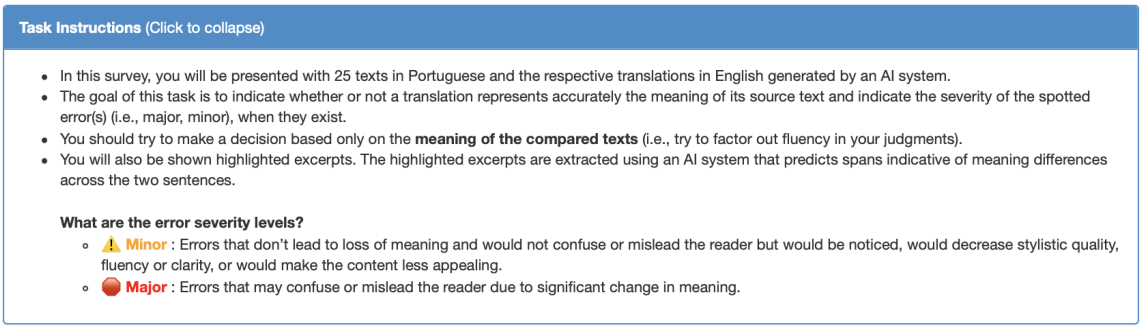


Figure 6.8: Instructions for application-grounded evaluation II: Critical Error Detection.

described in §6.4.

Procedures & Participants We collect 5 assessments per instance and annotate a total of 100 instances (50 per condition). Each participant is presented with 25 instances, among which 20% correspond to WMT critical errors, 20% to negation errors, and the rest to original OPUS parallel texts. We recruited a total of 40 participants, all of whom self-identified as proficient in English and Portuguese. All other participant recruitment details and study procedures are the same as in § 6.4. We include screenshots of the annotation task interface in Figure 6.10.

6.5.3 Study Results

Main Findings As shown in Figure 6.9, the user group presented with contrastive phrasal highlights exhibits a higher recall in detecting critical errors compared to the one that does not access them. Crucially, we report significant improvements ($p = 0.05$) in detecting *negation* errors, which are harder to spot. We note that although the improvements in detecting WMT errors are not-significant ($p > 0.05$) they remain significant overall (i.e., ALL set). A closer look at the differences between the two error classes reveals that detecting the former is more challenging as annotators have to pay close attention to the compared texts and the relationships between them to spot local mistranslations that cause a major shift in

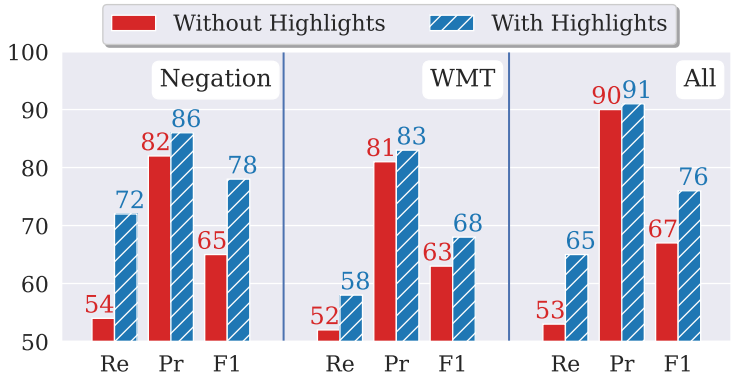
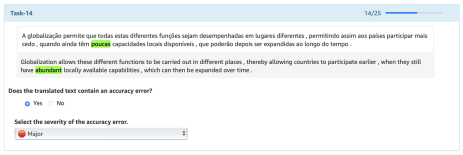


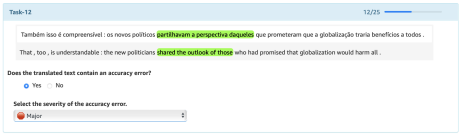
Figure 6.9: Application-grounded evaluation results comparing with vs. without highlights. Contrastive phrasal highlights significantly improve the Recall and F1 scores when detecting *negation* errors ($p = 0.05$) and errors in the ALL set. Precision improvements across sets and any improvements on the WMT set are not significant ($p > 0.05$).

meaning. On the other hand, WMT errors are more easily spotted as they mostly represent detached hallucination phenomena (i.e., an entire phrase has been added to the English translation), which we hypothesize can be implied by length differences of the compared texts. Examples of annotations are included in Figure 6.10.

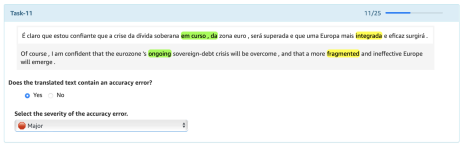
Subjective Measures & User Feedback Bilingual speakers found the highlights helpful overall, with an average self-reported usefulness of 3.9. Additionally, they reported they would like to use them as a tool to assist them in detecting critical errors, with an average score of 3.8. A closer look at the user's feedback sheds some light on the current strengths and limitations of our approach. Although highlights were in principle useful and a *"good feature to have especially when proofreading a great amount of translated texts"*, some annotators spotted *"a high percentage of false positives"* (see examples 6.10i and 6.10j) and raised concerns that *"relying solely on them might make them less aware of errors hidden in non-highlighted texts"* (see example 6.10b).



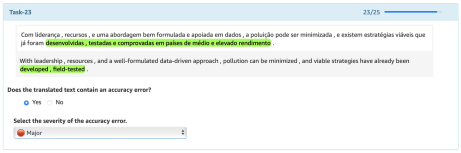
(a) **Major error** (negation)



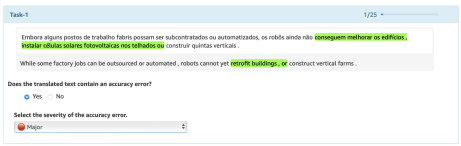
(b) **Major error** (negation)



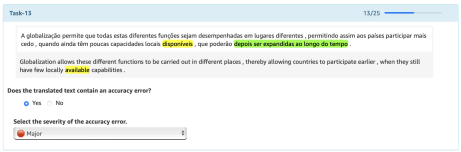
(c) **Major error** (negation)



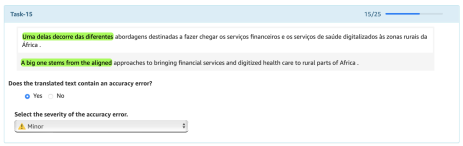
(d) **Major error** (hallucination)



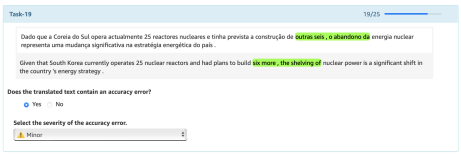
(e) **Major error** (hallucination)



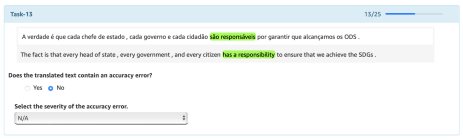
(f) **Major error** (hallucination)



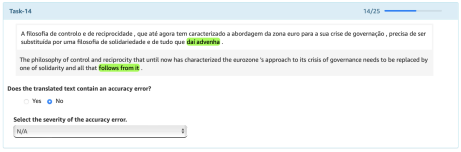
(g) **Minor error**



(h) **Minor error**



(i) **No error**



(j) **No error**

Figure 6.10: Annotations of accuracy errors in Portuguese translations of English texts. Contrastive highlights frequently surface meaning differences across the compared texts.

6.6 Summary

Our last piece of work explored ways of re-framing divergence prediction in ways that can be directly useful to humans. In doing so, we introduced a lightweight approach to extracting contrastive phrasal highlights that explain cross-lingual semantic divergences by considering the input's structure and explicitly modeling the relationships between contrasting inputs. Our findings suggest that contrastive phrasal highlights match human-provided rationales of divergences in English-French Wikipedia texts better than standardized highlighting approaches. Finally and more importantly, we showed that contrastive phrasal highlights can assist humans in detecting fine-grained meaning differences in (human) translated texts and critical errors due to local mistranslations in machine-translated texts.

In the following chapter, we ask: ***What kinds of future work does this thesis enable?*** We initiated the discussion around this question by summarizing the thesis findings and contributions and reflecting on the limitations and unanswered questions that arose throughout our exploration.

Chapter 7: Conclusion and Future Work

7.1 Summary

This dissertation contributes data, methods, and algorithms that together **advance our *understanding* of machine and human translation** through the lens of fine-grained semantic divergences. Throughout this thesis, we have been questioning the idea that a source text and its translation are equivalent in meaning—a common hypothesis that drives the way we think about models in machine translation research and the way humans consume translations.

Our first work contributed protocols, data, and models that improved both the annotation and prediction of fine-grained semantic divergences providing the foundation for studying their connection to machine translation (Briakou and Carpuat, 2020). In this direction, we introduced the Rationalized English-French Semantic Divergences corpus, based on a novel divergence annotation protocol that asks for human rationales to improve annotator agreement. We provided empirical evidence showing that semantic divergences are surprisingly frequent (40%) in WikiMatrix—a large-scale corpus routinely used in machine translation. Additionally, drawing on translation studies, we introduced computational methods for detecting them accurately at scale. Crucially, although our findings are limited to a single language-pair, i.e., English-French, and one domain, i.e., Wikipedia, our approach *does not* assume access to human-annotated samples of divergence, opening up the possibility of studying divergences across more languages and domains than covered in this dissertation.

In our second piece of work, we contributed empirical evidence and models for improving machine translation understanding through the lens of semantic divergences. We showed that those small divergences hurt the translation accuracy and confidence of neural machine translation models and crucially are one of the root causes that connect to known pathologies of neural text generation (Briakou and Carpuat, 2021). After establishing that divergences matter for machine translation training, we explored different approaches to mitigating their impact across a diverse range of languages and data regimes. First, we introduced a divergent-aware neural translation system (Briakou and Carpuat, 2021) and showed that modeling divergences explicitly helps mitigate their negative impact in a high-resource setting. Then, we introduced different editing models better suited to medium (Briakou and Carpuat, 2022b) and low resource regimes (Briakou et al., 2022), improving machine translation quality in 7 language pairs (paired with English) and 14 directions. Although our exploration of machine translation has covered a wide range of languages, covering different domains, language families, and data sizes, we have almost solely evaluated our proposed approach on English-centric contexts, where we expect our methods to perform better. Therefore, our work raises the question of whether divergences can further drive research in extremely low-resource and non-English-centric contexts, where filtering out data is suboptimal, and in different translation paradigms where parallel texts provide an implicit signal of supervision.

In our last piece of work, we introduced an approach to automatically predict not just *whether* semantic divergences exist in bilingual texts but also *where* they reside. Drawing on social science studies, we introduced a method to extract contrastive phrasal highlights that explain the predictions of our divergent detectors by explicitly modeling the relationships between the contrasted texts. We contributed evidence that contrastive phrasal highlights match human-provided rationales of divergence better than standard highlighting approaches, and more importantly, they assist bilingual speakers in annotating fine-grained divergences—

easing the need to ask for human rationales. Finally, we showed that contrastive highlights could help humans detect critical errors due to local mistranslations in machine-translated texts. Although contrastive highlights have been found helpful in our user studies, how our method performs when it comes to detecting translation differences in the wild (which requires covering the entire distribution of meaning differences), as well as in other language settings, remain open questions. Nevertheless, our findings create opportunities for designing better machine-in-the-loop pipelines to identify critical machine translation errors grounded in high-stake settings, study translation data at scale, and facilitate the creation of multilingual content through crowd-based efforts.

7.2 Future Work

In this section, drawing on the findings and contributions of this thesis, we we describe several avenues for future research to improve translation and multilingual NLP understanding further.

7.2.1 Semantic Divergences in the Era of Large Language Models

This thesis has examined how fine-grained semantic divergences impact supervised neural machine translation systems assuming that parallel corpora are a crucial component for training translation systems by means of direct supervision. While this is still true to a large extent, recent work with large language models shows that parallel texts are used very differently than in their supervised counterparts. Concretely, we know that parallel texts are used in (at least) two ways: they provide an implicit training signal as they are incidentally consumed within the large amounts of monolingual data as a byproduct of scale (Briakou et al., 2023) or an explicit in-context learning signal as a contextualized input following the few-shot learning prompting paradigm (Vilar et al., 2022; Agrawal

et al., 2022). As a result, understanding the role of fine-grained divergences in the era of large language models opens up new research questions. How do semantic divergences that are incidentally consumed in the pre-training data impact machine translation capabilities? Can they explain hallucination or other degeneration phenomena, similarly to how they impact supervised NMT systems? Do they connect with the large language models' abilities to produce less literal translations (Raunak et al., 2023)? How can we use translation divergences as in-context learning demonstrations to tailor machine translation to mimic (desired) divergences?

7.2.2 Beyond Highlighting Semantic Divergences

This thesis has provided evidence that contrastive phrasal highlights can provide a framework that assists humans in detecting meaning differences of a specific nature (e.g., fine-grained translation processes or local critical errors). However, detecting translation differences in the wild, requires covering the entire distribution of meaning differences (Freitag et al., 2021a). Therefore, what is the best way of assisting humans in such cases might require more advanced frameworks that move beyond highlights to verbalizing and characterizing the differences between texts more explicitly. As a result, more research should be conducted to understand what makes a good explanation for describing translation differences in concrete use cases. Furthermore, what constitutes a good explanation is a question that depends on a user's preference and knowledge. For instance, explaining translation differences to bilingual speakers, as done in this thesis, potentially requires different approaches than explaining translation differences to monolingual speakers who have limited or no knowledge of the source language and potentially could benefit from additional background knowledge, i.e., from translation lexicons. Those scenarios give rise to a series of questions: Should translation differences be described by surfacing how the

segments contrast and compare in *language form* instead of highlighting them? How can we adapt explanations based on users' preferences and language proficiency? How do different approaches to explaining cross-lingual meaning differences address issues of overreliance?

7.2.3 Improving Multilingual NLP Through the lens of Semantic Divergences

This thesis used semantic divergences as a tool to understand machine translation; however, the former can provide a framework for understanding and analyzing multilingual NLP more broadly. For instance, a scalable way to build multilingual resources is to use translation to port data labeled in English to other languages (Conneau et al., 2018; Yang et al., 2019; Ruder et al., 2021). Yet such translation-based data can be problematic in different ways ranging from errors introduced through translation—to the representation of linguistic or cultural concepts that do not accurately reflect the real-world use of those languages. Furthermore, building annotated datasets from scratch for each language does not scale easily (Briakou et al., 2021). Can we have the best of both worlds? How can we build better multilingual benchmarks by marrying the scalable strengths of generative AI, the interpretable strengths of explainable semantic divergences, and the evaluative strengths of *crowdworkers*?

Bibliography

Andira Abdallah. 2021. Impact of using parallel text strategy on teaching reading to intermediate ii level students.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Neural Information Processing Systems*.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main*

Conference and the Shared Task: Semantic Textual Similarity, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *ArXiv*, abs/2212.02437.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7(0):597–610.

Tom Ash, Remi Francis, and Will Williams. 2018. The speechmatics parallel corpus filtering system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 853–859, Belgium, Brussels. Association for Computational Linguistics.

Alhassan Awad, Sabtan Yasser, Muhammed Naguib Sabtan, and Omar Lamis. 2021. Using parallel corpora in the translation classroom: Moving towards a corpus-driven pedagogy for omani translation major students.

Andoni Azpeitia and Thierry Etchegoyhen. 2016. DOCAL - vicomtech's participation in the WMT16 shared task on bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 666–671, Berlin, Germany. Association for Computational Linguistics.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: Bridging the Wikipedia Language Gap. In *Proceedings*

of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12, pages 1075–1084, Austin, Texas, USA. ACM.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the multisemcor corpus. *Natural Language Engineering*, 11:247 – 261.

Rahul Bhagat and Eduard Hovy. 2013. What Is a Paraphrase? *Computational Linguistics*, 39(3):463–472.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland. Association for Computational Linguistics.

Brody Bluemel. 2014. Learning in parallel: Using parallel corpora to enhance written language acquisition at the beginning level.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Q. Vera Liao, Tania Lombrozo, Alison Smith-Renner, and Chenhao Tan. 2022. Human-centered evaluation of explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 26–32, Seattle, United States. Association for Computational Linguistics.

Fabienne Braune and Alexander Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Coling 2010: Posters*, pages 81–89, Beijing, China. Coling 2010 Organizing Committee.

Eleftheria Briakou and Marine Carpuat. 2020. Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1563–1580, Online. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2021. Beyond noise: Mitigating the impact of fine-grained semantic divergences on neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7236–7249, Online. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2022a. Can synthetic translations improve bitext quality? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (to appear)*, Online. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2022b. Can synthetic translations improve bitext quality? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4753–4766, Dublin, Ireland. Association for Computational Linguistics.

Eleftheria Briakou, Colin Cherry, and George F. Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in palm's translation capability. *ArXiv*, abs/2305.10266.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Eleftheria Briakou, Sida Wang, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. BitextEdit: Automatic bitext editing for improved low-resource machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1469–1485, Seattle, United States. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176, Berkeley, California, USA. Association for Computational Linguistics.

Zana Buccinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces*.

Christian Buck and Philipp Koehn. 2016a. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Christian Buck and Philipp Koehn. 2016b. Quick and reliable document alignment via TF/IDF-weighted cosine distance. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 672–678, Berlin, Germany. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Neural Information Processing Systems*.

Rémi Cardon and Natalia Grabar. 2020. Reducing the search space for parallel sentences in comparable corpora. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 44–48, Marseille, France. European Language Resources Association.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Vancouver. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT evaluation campaign. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Evaluation Campaign*, Heidelberg, Germany.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 2–17, Lake Tahoe, California.

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the WMT 2019 shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28, Florence, Italy. Association for Computational Linguistics.

Rajen Chatterjee, Markus Freitag, Matteo Negri, and Marco Turchi. 2020. Findings of the WMT 2020 shared task on automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 646–659, Online. Association for Computational Linguistics.

Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018. Findings of the WMT 2018 shared task on automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 710–725, Belgium, Brussels. Association for Computational Linguistics.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. Association for Computational Linguistics.

Saneem Chemmengath, Amar Prakash Azad, Ronny Luss, and Amit Dhurandhar. 2022. Let the CAT out of the bag: Contrastive attributed explanations for text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7190–7206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qianglong Chen, Feng Ji, Xiangji Zeng, Feng-Lin Li, Ji Zhang, Haiqing Chen, and Yin Zhang. 2021. KACE: Generating knowledge aware contrastive explanations for natural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2516–2527, Online. Association for Computational Linguistics.

Stanley F. Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Columbus, Ohio, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. *Cross-Lingual Language Model Pretraining*. Curran Associates Inc., Red Hook, NY, USA.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, Michigan. Association for Computational Linguistics.

Josep Maria Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurélien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. 2016. Systran's pure neural machine translation systems. *CoRR*, abs/1610.05540.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

Aswarth Abhilash Dara and Yiu-Chang Lin. 2016. YODA system for WMT16 shared task: Bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 679–684, Berlin, Germany. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Rajani, Eric P. Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. In *Annual Meeting of the Association for Computational Linguistics*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020a. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020b. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 255–262, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING*.

Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. The sockeye 2 neural machine translation toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 110–115, Virtual. Association for Machine Translation in the Americas.

Bonnie J. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *CL*, 20(4).

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

K. Durán, J. Rodríguez, and M. Bravo. 2014. Similarity of sentences through comparison of syntactic trees with pairs of similar words. In *2014 11th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, pages 1–6.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Miquel Esplà-Gomis. 2009. Bitextor: a free/open-source software to harvest translation memories from multilingual websites. In *Beyond Translation Memories: New Tools for Translators Workshop*, Ottawa, Canada.

Miquel Esplà-Gomis, Mikel Forcada, Sergio Ortiz-Rojas, and Jorge Ferrández-Tordera. 2016. Bitextor's participation in WMT'16: shared task on document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 685–691, Berlin, Germany. Association for Computational Linguistics.

Allyson Ettinger, Philip Resnik, and Marine Carpuat. 2016. Retrofitting sense-specific word vectors using parallel text. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1378–1383, San Diego, California. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Giorgos Floros. 2004. Parallel texts in translating and interpreting.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, George F. Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021b. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021c. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Pascale Fung and Percy Cheung. 2004a. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and E. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 57–63, Barcelona, Spain. Association for Computational Linguistics.

Pascale Fung and Percy Cheung. 2004b. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1051–1057, Geneva, Switzerland. COLING.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2016. Factored neural machine translation. *CoRR*, abs/1609.04621.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2017. Neural machine translation by generating multiple linguistic factors. *CoRR*, abs/1712.01821.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. 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 4453–4462, Hong Kong, China. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for datasets. *ArXiv*, abs/1803.09010.

Luís Gomes and Gabriel Pereira Lopes. 2016. First steps towards coverage-based sentence alignment. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2228–2231, Portorož, Slovenia. European Language Resources Association (ELRA).

Luís Gomes and Gabriel Pereira Lopes. 2016. First steps towards coverage-based document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 697–702, Berlin, Germany. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, International Convention Centre, Sydney, Australia. PMLR.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Barry Haddow and Philipp Koehn. 2012. Interpolated backoff for factored translation models. In *Association for Machine Translation in the Americas, AMTA*.

Zellig Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1576–1586, Lisbon, Portugal. Association for Computational Linguistics.

Denis Hilton. 2017. Social attribution and explanation. *Oxford Handbook of Causal Reasoning*, page 645–676.

Graeme Hirst. 1995. Near-synonymy and the structure of lexical knowledge. In *AAAI Symposium on Rep- resentation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity. pages 51– 56.*, pages 51–56.

Cong Duy Vu Hoang, Gholamreza Haffari, and Trevor Cohn. 2016. Improving neural translation models with linguistic factors. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 7–14, Melbourne, Australia.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara I. Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11:311 – 325.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *ACL*.

Alon Jacovi and Yoav Goldberg. 2021. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *North American Chapter of the Association for Computational Linguistics*.

Laurent Jakubina and Phillippe Langlais. 2016. BAD LUC@WMT 2016: a bilingual document alignment platform based on lucene. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 703–709, Berlin, Germany. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Wenxiang Jiao, Xing Wang, Shilin He, Irwin King, Michael Lyu, and Zhaopeng Tu. 2020. Data Rejuvenation: Exploiting Inactive Training Examples for Neural Machine Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2255–2266, Online. Association for Computational Linguistics.

Isaac Johnson and Emily A. Lescak. 2022. Considerations for multilingual wikipedia research. *ArXiv*, abs/2204.02483.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. *CoRR*, abs/1610.01108.

Mihir Kale, Aditya Siddhant, Rami Al-Rfou, Linting Xue, Noah Constant, and Melvin Johnson. 2021. nmT5 - is parallel data still relevant for pre-training massively multilingual language models? In *Proceedings of the 59th Annual Meeting of the Association*

for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 683–691, Online. Association for Computational Linguistics.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. 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 3336–3341, Hong Kong, China. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2019a. Multi-task stack propagation for neural quality estimation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(4).

Hyun Kim, Joon-Ho Lim, Hyun-Ki Kim, and Seung-Hoon Na. 2019b. QE BERT: Bilingual BERT using multi-task learning for neural quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 85–89, Florence, Italy. Association for Computational Linguistics.

Siwon Kim, Jihun Yi, Eunji Kim, and Sungroh Yoon. 2020. Interpretation of NLP models through input marginalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3154–3167, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In

Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pages 54–72, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022a. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman,

Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022b. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Thanh C. Le, Hoa Trong Vu, Jonathan Oberländer, and Ondřej Bojar. 2016. Using term position similarity and language modeling for bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 710–716, Berlin, Germany. Association for Computational Linguistics.

Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the World Reads Wikipedia: Beyond English Speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pages 618–626, New York, NY, USA. ACM.

Chuanrong Li, Lin Shengshuo, Zeyu Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. 2020. Linguistically-informed transformations (LIT): A method for automatically generating contrast sets. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 126–135, Online. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *ArXiv*, abs/1612.08220.

Zhongyang Li, Tongfei Chen, and Benjamin Van Durme. 2019. Learning to rank for plausible plausibility. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4818–4823, Florence, Italy. Association for Computational Linguistics.

Qingzi Vera Liao, Dan Gruen, and Sarah Miller. 2020. Questioning the ai: Informing design practices for explainable ai user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Chi-kiu Lo, Michel Simard, Darlene Stewart, Samuel Larkin, Cyril Goutte, and Patrick Littell. 2018. Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The NRC supervised submissions to the parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 908–916, Belgium, Brussels. Association for Computational Linguistics.

Pintu Lohar, Haithem Afli, Chao-Hong Liu, and Andy Way. 2016. The ADAPT bilingual document alignment system at WMT16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 717–723, Berlin, Germany. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. VECO: Variable and flexible cross-lingual pre-training for language understanding and generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3980–3994, Online. Association for Computational Linguistics.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands. Association for Computational Linguistics.

Sainik Mahata, Dipankar Das, and Santanu Pal. 2016. WMT2016: A hybrid approach to bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 724–727, Berlin, Germany. Association for Computational Linguistics.

Paolo Massa and Federico Scrinzi. 2012. Manypedia: Comparing Language Points of View of Wikipedia Communities. In *Proceedings of the Eighth Annual International*

Symposium on Wikis and Open Collaboration, WikiSym '12, pages 21:1–21:9, Linz, Austria. ACM.

Marek Medveď, Miloš Jakubíček, and Vojtech Kovář. 2016. English-French document alignment based on keywords and statistical translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 728–732, Berlin, Germany. Association for Computational Linguistics.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sc iences. *Artificial Intelligence*, 267:1–38.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 135–144, Tiburon, USA. Springer.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia. Association for Computational Linguistics.

M. Lynne Murphy. 2010. *Lexical Meaning*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2012. Semeval-2012 task 8: Cross-lingual textual entailment for content synchronization. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 399–407, Montréal, Canada. Association for Computational Linguistics.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. ESCAPE: a large-scale synthetic corpus for automatic post-editing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. *Proceedings of the AAAI Conference on Artificial Intellige nce*, 34(05):8568–8575.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965. PMLR.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *J. Artif. Int. Res.*, 36(1):307–340.

Vassilis Papavassiliou, Prokopis Prokopidis, and Stelios Piperidis. 2016. The ILSP/ARC submission to the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 733–739, Berlin, Germany. Association for Computational Linguistics.

Vassilis Papavassiliou, Sokratis Sofianopoulos, Prokopis Prokopidis, and Stelios Piperidis. 2018. The ILSP/ARC submission to the WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 928–933, Belgium, Brussels. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, Online. Association for Computational Linguistics.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. 2015. Adding semantics to data-driven paraphrasing. 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 1512–1522, Beijing, China. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jan-Thorsten Peter, Arne Nix, and Hermann Ney. 2017. Generating alignments using target foresight in attention-based neural machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108:27 – 36.

MinhQuang Pham, Josep Crego, Jean Senellart, and François Yvon. 2018. Fixing translation divergences in parallel corpora for neural MT. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2967–2973, Brussels, Belgium. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Jipeng Qiang, Yifan Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2019. A simple bert-based approach for lexical simplification. *ArXiv*, abs/1907.06226.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan Awadallah. 2023. Do gpts produce less literal translations? *ArXiv*, abs/2305.16806.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philip Resnik. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527–534, College Park, Maryland, USA. Association for Computational Linguistics.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. *CoRR*, abs/1509.06664.

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online. Association for Computational Linguistics.

Nick Rossenbach, Jan Rosendahl, Yunsu Kim, Miguel Graça, Aman Gokrani, and Hermann Ney. 2018. The RWTH Aachen University filtering system for the WMT 2018 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 946–954, Belgium, Brussels. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Vadim Shchukin, Dmitry Khristich, and Irina Galinskaya. 2016. Word clustering approach to bilingual document alignment (WMT 2016 shared task). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 740–744, Berlin, Germany. Association for Computational Linguistics.

Haoyue Shi, Luke Zettlemoyer, and Sida I. Wang. 2021. Bilingual lexicon induction via unsupervised bitext construction and word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 813–826, Online. Association for Computational Linguistics.

Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. 2006. A DOM tree alignment model for mining parallel data from the web. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 489–496, Sydney, Australia. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *ArXiv*, abs/1706.03825.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. 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.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Kaiser Sun and Ana Marasović. 2021. Effective attention sheds light on interpretability. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4126–4135, Online. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*.

Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. 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 1556–1566, Beijing, China. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Jörg Tiedemann. 2012a. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jrg Tiedemann. 2011. *Bitext Alignment*, 1st edition. Morgan and Claypool Publishers.

Jörg Tiedemann. 2012b. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Marcos Treviso, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2021. IST-unbabel 2021 submission for the explainable quality estimation shared task. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 133–145, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Tutek and Jan Snajder. 2020. Staying true to your word: (how) can attention become explanation? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 131–142, Online. Association for Computational Linguistics.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1101–1109, Beijing, China. Coling 2010 Organizing Committee.

Dániel Varga, Péter Halácsy, András Kornai, Nagy Viktor, Nagy Laszlo, Németh László, and Tron Viktor. 2007. Parallel corpora for medium density languages.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llio n Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George F. Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *ArXiv*, abs/2211.09102.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018a. Identifying semantic divergences in parallel text without annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1503–1515, New Orleans, Louisiana. Association for Computational Linguistics.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018b. Identifying semantic divergences in parallel text without annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1503–1515, New Orleans, Louisiana. Association for Computational Linguistics.

Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020a. Gradient-based analysis of NLP models is manipulable. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 247–258, Online. Association for Computational Linguistics.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020b. On the inference calibration of neural machine translation. In *ACL*.

Shira Wein, Lucia Donatelli, Ethan Ricker, Calvin Engstrom, Alex Nelson, Leonie Harter, and Nathan Schneider. 2022. Spanish Abstract Meaning Representation: Annotation of a general corpus. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

Shira Wein and Nathan Schneider. 2021. Classifying divergences in cross-lingual AMR pairs. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 56–65, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Dekai Wu and Pascale Fung. 2005. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Second International Joint Conference on Natural Language Processing: Full Papers*.

Yonghui Wu, Mike Schuster, Z. 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 R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. 2016. Growing Wikipedia Across Languages via Recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 975–985, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Weijia Xu, Shuming Ma, Dongdong Zhang, and Marine Carpuat. 2021. How does distilled data complexity impact the quality and confidence of non-autoregressive machine translation? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4392–4400, Online. Association for Computational Linguistics.

Wenda Xu, Yi-Lin Tuan, Yujie Lu, Michael Saxon, Lei Li, and William Yang Wang. 2022. Not all errors are equal: Learning text generation metrics using stratified error synthesis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6559–6574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Ching-Man Au Yeung, Kevin Duh, and Masaaki Nagata. 2011. Providing Cross-Lingual Editing Assistance to Wikipedia Editors. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'11, pages 377–389, Tokyo, Japan. Springer-Verlag.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007a. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Omar F. Zaidan, Jason Eisner, and Christine Piatko. 2007b. Using "annotator rationales" to improve machine learning for text categorization. In *NAACL HLT 2007; Proceedings of the Main Conference*, pages 260–267.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2019. Adding interpretable attention to neural translation models improves word alignment. *ArXiv*, abs/1901.11359.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yuming Zhai, Aurélien Max, and Anne Vilnat. 2018. Construction of a multilingual corpus annotated with translation relations. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 102–111, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yuming Zhai, Pooyan Safari, Gabriel Illouz, A. Allauzen, and Anne Vilnat. 2019a. Towards recognizing phrase translation processes: Experiments on english-french. *POLIBITS*, 60:27–33.

Yuming Zhai, Pooyan Safari, Gabriel Illouz, Alexandre Allauzen, and Anne Vilnat. 2019b. Towards recognizing phrase translation processes: Experiments on english-french. *CoRR*, abs/1904.12213.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Chunting Zhou, Graham Neubig, and Jiatao Gu. 2019a. Understanding knowledge distillation in non-autoregressive machine translation. *CoRR*, abs/1911.02727.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019b. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.