ABSTRACT

Title of dissertation: Identifying Semantic Divergences Across Languages

Yogarshi Vyas
Doctor of Philosophy, 2019

Dissertation directed by: Professor Marine Carpuat
Department of Computer Science

Cross-lingual resources such as parallel corpora and bilingual dictionaries are cornerstones of multilingual natural language processing (NLP). They have been used to study the nature of translation, train automatic machine translation systems, as well as to transfer models across languages for an array of NLP tasks. However, the majority of work in cross-lingual and multilingual NLP assumes that translations recorded in these resources are semantically equivalent. This is often not the case—words and sentences that are considered to be translations of each other frequently *diverge* in meaning, often in systematic ways.

In this thesis, we focus on such mismatches in meaning in text that we expect to be aligned across languages. We term such mismatches as *cross-lingual semantic divergences*. The core claim of this thesis is that translation is not always meaning preserving which leads to cross-lingual semantic divergences that affect multilingual NLP tasks. Detecting such divergences requires ways of directly characterizing differences in meaning across languages through novel cross-lingual tasks, as well

as models that account for translation ambiguity and do not rely on expensive, task-specific supervision.

We support this claim through three main contributions. First, we show that a large fraction of data in multilingual resources (such as parallel corpora and bilingual dictionaries) is identified as semantically divergent by human annotators. Second, we introduce cross-lingual tasks that characterize differences in word meaning across languages by identifying the semantic relation between two words. We also develop methods to predict such semantic relations, as well as a model to predict whether sentences in different languages have the same meaning. Finally, we demonstrate the impact of divergences by applying the methods developed in the previous sections to two downstream tasks. We first show that our model for identifying semantic relations between words helps in separating equivalent word translations from divergent translations in the context of bilingual dictionary induction, even when the two words are close in meaning. We also show that identifying and filtering semantic divergences in parallel data helps in training a neural machine translation system twice as fast without sacrificing quality.

Identifying Semantic Divergences Across Languages

by

Yogarshi Paritosh Vyas

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
2019

Advisory Committee:
Professor Marine Carpuat, Chair/Advisor
Professor Jordan Boyd-Graber
Professor Ido Dagan
Professor David Jacobs
Professor Philip Resnik (Dean's Representative)

Acknowledgments

I cannot overstate the contribution of my advisor, Marine Carpuat, to this entire process. I learned many things from Marine, all of which are impossible to summarize in a few sentences, but everything she taught me has contributed to making me a better researcher. Working with Marine was like combining the best of many different worlds. She allowed me immense flexibility to choose research problems, but at the same time, was readily available to provide sharp technical insights in the face of roadblocks. Marine also taught me the value of clear thinking and exposition, and constantly motivated me to write better papers, and give better talks. I hope to have inculcated a small fraction of all these things.

I wish to thank the members of my dissertation committee. Philip Resnik's seminar on semantics in my very first semester was instrumental in my earliest projects, and revealed many hitherto unknown avenues of computational linguistics and semantics. It is safe to say that many parts of this thesis would not exist without the foundational work of Ido Dagan and his collaborators. Jordan Boyd-Graber and David Jacobs asked the right questions when this thesis was proposed, and I am grateful for all the feedback, then and now.

During my time at Maryland, I have had the opportunity to work with many members of CLIP Lab, and I have learned many things from my interactions with these talented people over the years. I would like to thank Hal Daumé III and Sudha Rao, who along with Philip formed my first set of collaborators in the CLIP lab. Sudha was extremely helpful in showing me the ropes of being a grad student

in my early grad school days. I could have never imagined I would be part of an academic paper on comic books, but that is a testament to the boundless creativity of Mohit Iyyer, Varun Manjunatha, and Anupam Guha. Thanks for being a part of the most fun project I've done in grad school, and also thanks for many hours of fun and camaraderie, both within and outside the lab. Thanks also to Joe Barrow, Pedro Rodriguez, and Han-Chin Shing for the countless conversations while waiting for models to converge, and bubble tea walks, and board game nights. Also thanks to Xing Niu, Marianna Martindale, Weijia Xu, Sweta Agrawal, Ahmed Elgohary, Amittai Axelrod, Allyson Ettinger, Ahmed Elgohary, Suraj Nair, Jeff Green, Rachel Adler, Naomi Feldman, and all other members of CLIP and the Linguistics department for countless stimulating conversations and feedbacks on paper drafts.

My interaction with colleagues outside of UMD have provided me diverse perspectives on research. I spent many months working with Shyam Upadhyay on the cross-lingual hypernymy project, and despite the constant roadblocks and dead-ends, the collaboration was very a fruitful experience. Jags (Jagadeesh Jagarlamudi) was a great mentor during my summer at Google. Georgiana Dinu, Rishita Anubhai, Jie Ma, and Yaser Al-Onaizan have been very friendly mentors and collaborators during my summer at Amazon and I continue to learn from them and grow.

Friends and family have brought balance to my life and served as buoys in a sea of academic turbulence. Karthik, Manasij, and Janhavi have managed to tolerate me for countless hours through ups and downs. They along with Jay, Jaideep, and Nidhi were fun roommates and steadfast friends, and the countless hours we spent hanging out in Catawba and GH were the comfort food equivalent of my social

interactions. Having old friends who anchor you in new locations is a luxury, and Abhishek has done exactly that over the two summers I spent in NYC.

Family has also been a source of patient and constant support (only occasionally punctuated by the dreaded "*When will you graduate?*" question). My parents, Maitri and Paritosh, and my sister, Riddhi have been a constant source of encouragement and love, and my India trips to visit them have been a mental reset button. A special shout-out (meow?) to Shadow, my furry friend and feline companion, who has been a literal stressbuster these last few months. And of course, Neha, for not only all the support and love, but also for making sure I take breaks, breathe, and find the healthy balance between academia and life beyond. I cannot be more grateful.

Table of Contents

Acknowledgements ii

Table of Contents v

List of Tables viii

List of Figures xi

1 Introduction 1
- 1.1 Thesis Statement . . . 4
- 1.2 Roadmap . . . 4
- 1.3 Contributions . . . 7

2 Background 9
- 2.1 Translation . . . 9
  - 2.1.1 Meaning and Semantic Equivalence . . . 9
  - 2.1.2 Translation Equivalence is not Semantic Equivalence . . . 11
  - 2.1.3 Automatically Identifying Translation Correspondences using Alignment . . . 13
  - 2.1.4 Semantic Divergences in Noisy Parallel Data . . . 14
  - 2.1.5 Translation Divergences . . . 16
- 2.2 Divergences in Sentences Beyond Parallel Data . . . 17
- 2.3 Lexical Divergences and Semantic Relations . . . 21
  - 2.3.1 Non-equivalence of Lexical Translations . . . 21
  - 2.3.2 Lexico-semantic Relations . . . 24
  - 2.3.3 Automatic Methods for Detecting Hypernymy and Hyponymy 27
  - 2.3.4 Automatic Methods for Other Relations . . . 30
  - 2.3.5 Semantic Relations across Languages . . . 31
  - 2.3.6 Utility in downstream tasks . . . 32
- 2.4 Summary . . . 32

3 Quantifying and Analyzing Divergences in Bilingual Resources 34
- 3.1 Divergences in Bilingual Dictionaries . . . 34
- 3.2 Divergences in Parallel Sentences . . . 37
- 3.3 Summary . . . 39

4 Identifying Cross-lingual Hypernymy using Sparse Bilingual Embeddings 41
4.1 Cross-lingual Hypernymy : Challenges and Contributions 42
4.2 Unsupervised Identification of Cross-lingual Hypernymy 44
4.3 BiSparse-Dep: Sparse Bilingual Word Representations using Dependency Contexts 46
4.3.1 Review: Learning Monolingual Sparse Representations 47
4.3.2 BiSparse: Learning Sparse Bilingual Embeddings 48
4.3.3 BiSparse-Dep: Inducing Dependency Based Contexts 50
4.3.4 Dependency Contexts without a Treebank 51
4.3.5 Optimization of BiSparse formulation 52
4.4 Crowd-Sourcing Annotations 53
4.4.1 Annotation Setup 53
4.4.2 Two Evaluation Test Sets 55
4.5 Experimental Setup 56
4.5.1 Data and Evaluation Setup 56
4.5.2 Contrastive Approaches 58
4.5.3 Evaluating Robustness of Bisparse-Dep 59
4.6 Experiments 61
4.6.1 Dependency v/s Window Contexts 62
4.6.2 Ablating Directionality in Context 64
4.6.3 Evaluating Robustness of BiSparse-Dep 64
4.6.4 Choice of Entailment Scorer 67
4.7 Summary 68

5 Weakly Supervised Identification of Cross-lingual Semantic Relations 70
5.1 BiLexNet: a Classifier for Cross-Lingual Semantic Relations 72
5.1.1 Cross-lingual Paths 73
5.2 Weakly Supervised Training via Knowledge Distillation 75
5.3 MultiLexRel : A Dataset for Cross-lingual Semantic Relations . . 78
5.4 Experimental Settings 79
5.4.1 Data 79
5.4.2 Model Configurations and Baselines 81
5.5 Results 84
5.6 Analysis 86
5.7 A Comparison of Unsupervised and Weakly Supervised Approaches . 89
5.8 Summary 91

6 Identifying Semantic Divergences in Parallel Text without Annotations 93
6.1 Background 94
6.2 Approach 95
6.3 Divergence Detection Evaluation 98
6.3.1 Neural Semantic Divergence Detection 99
6.3.2 Parallel vs. Non-parallel Classifier 99
6.3.3 Neural MT 100
6.3.4 Bilingual Sentence Embeddings 101

6.3.5 Textual Entailment Classifier . . . 102
6.4 Intrinsic Evaluation Results . . . 103
6.4.1 Analysis . . . 105
6.5 Summary . . . 105

7 Divergences in NLP Applications 107
7.1 Semantic Divergences in Automatically Constructed Dictionaries . . . 108
7.1.1 Background: Bilingual Dictionary Induction via Bilingual Embeddings . . . 109
7.1.2 Using BiLexNet to Filter Divergent Lexical Translations . . . 110
7.1.3 Setup and Data . . . 112
7.1.4 Results and Discussion . . . 113
7.2 Improving Neural MT Training by Filtering Semantic Divergences . . . 114
7.2.1 Translation Tasks . . . 115
7.2.2 Neural MT System . . . 116
7.2.3 English-French Results . . . 117
7.2.4 Vietnamese-English Results . . . 118
7.2.5 Analysis . . . 119
7.3 Summary . . . 121

8 Conclusion 122
8.1 Summary of Contributions . . . 122
8.2 Limitations and Future Work . . . 125
8.2.1 Fine-grained Distinction of Divergences . . . 125
8.2.2 Integrating Lexical and Sentential Models by Identifying Semantic Relations in Context . . . 127
8.2.3 Impact of Divergences and Cross-lingual Semantic Relations on other Cross-lingual Tasks . . . 129

Bibliography 131

List of Tables

1.1 Parallel sentences exhibit semantic divergences, as can be seen in these examples (English sentence (en), French sentence (fr) and its gloss (gl)) observed in a random sample of OpenSubtitles and the newstest2012 test set from WMT evaluations. 3

2.1 A unified summary of taxonomic semantic relations, natural logic relations (MacCartney and Manning, 2007), and the re-definition of the natural logic relations for PPDB (Pavlick et al., 2015), along with examples. 26

3.1 Examples of Hindi-English pairs labeled as divergent by annotators. . 37
3.2 Randomly selected sentence pairs (English (en), French (fr) and gloss of French (gl)) annotated as divergent or equivalent, with high and low degrees of agreement between the 5 annotators. Examples are taken from the OpenSubtitles (subs) and Common Crawl (cc) corpora. 40

4.1 Statistics of our crowd-sourced test bed. #pos (#neg) denote positives (negatives) in the evaluation set. We down-sample negatives to have a balanced evaluation set. 55
4.2 Training data statistics for different languages. Note that while we use parallel corpora for computing translation dictionaries, our approach does not require it, and can work with any bilingual dictionary. . . . 58
4.3 Comparing the different approaches from Section 4.5.2 with our BiSparse-Dep approach on Hyper-Hypo (random baseline = 0.5). **Bold** denotes the best score for each language, and the * on the best score indicates a statistically significant ($p < 0.05$) improvement over the next best score, using McNemar's test (McNemar, 1947). Dependency based models largely outperform window based models, and all BiSparse-Dep models outperform translation baselines. 61

4.4 Comparing the different approaches from Section 4.5.2 with our BiSparse-Dep approach on Hyper-Cohypo (random baseline = 0.5). **Bold** denotes the best score for each language, and the * on the best score indicates a statistically significant ($p < 0.05$) improvement over the next best score, using McNemar's test (McNemar, 1947). BiSparse-Dep models continue to outperform window based models and the translation baseline on an average. 62
4.5 The delexicalized model is competitive with the best dependency based and the best window based models on both test sets. For each dataset, * indicates a statistically significant ($p < 0.05$) improvement over the next best model in that column, using McNemar's test (McNemar, 1947). 65

5.1 Distribution of the five semantic relations for the two crowdsourced test sets. 79
5.2 Precision (P), Recall (R) and F1-score (F) for BiLexNet and contrastive baselines on the two MultiLexRel test sets. All configurations are trained with five random seeds. We report the mean score and standard deviation. The full BiLexNet model performs best and is consistently better with the attention component. 84
5.3 Per-class F1 scores for median En-Hi and En-Zh BiLexNet model and the EnLexNET model. 87
5.4 Accuracies of the weakly supervised BiLexNet when compared to the unsupervised BiSparse method from Chapter 4 on the task of cross-lingual hypernymy detection. The weak supervison enables BiLexNet to more accurately identify cross-lingual hypernymy than the unsupervised approach. 90

6.1 Intrinsic evaluation on crowdsourced semantic equivalence vs. divergence testsets. We report overall F-score, as well as precision (P), recall (R) and F-score (F) for the equivalent (+) and divergent (-) classes separately. Semantic similarity more accurately identifies divergences across the board, with larger improvements on the divergent class. 104

7.1 Combining predictions of semantic relations from BiLexNet with embeddings-based cosine similarity consistently improves Precision@1 for bilingual dictionary induction. Results in **bold** are statistically significant compared to next best result in the column (McNemar's test, $p < 0.05$). 112

7.2 English-French decoding results. BLEU scores are either averaged across 3 runs ("Avgerage") or obtained via ensemble decoding ("Ensemble"). SEMANTIC SIMILARITY reach BLEU scores on par with ALL with only half of the training data. SEMANTIC SIMILARITY scores marked with * are significanly better ($p < 0.05$) than the corresponding ALL scores. 117
7.3 Vietnamese-English decoding results: dropping 10% of the data based on SEMANTIC SIMILARITY does not hurt BLEU, and performs significantly ($p < 0.05$) better than RANDOM selection. 119
7.4 Selected translation examples from the ensemble systems of the various models. 120

List of Figures

2.1 Example of a WordNet synset (*gym shoe*, *sneaker*, *tennis shoe*), along with its hypernym (*shoe*), and its hyponym (*plimsoll*) synsets 27

3.1 Breakdown of a random sample of 500 examples from a Hindi-English bilingual dictionary with respect to annotator agreement. 36

4.1 The BiSparse-Dep approach, which learns sparse bilingual embeddings using dependency based contexts. The resulting sparse embeddings, together with an unsupervised entailment scorer, can detect hypernyms across languages (e.g., *pomme* is a *fruit*). 47
4.2 Example Dependency Tree. 51
4.3 BiSparse-Dep outperforms the best window based model on Hyper-Hypo, even with about 40% of the monolingual corpora, for most languages. 66
4.4 BiSparse-Dep outperforms the best window based model on Hyper-Hypo with increasingly lower quality dictionaries, for most languages. 67

5.1 On the left, we illustrate cross-lingual semantic relation classification: given the pair (*water*, **पय**) as input, the task is to select the `Equivalence` class (in bold/green) from the five possible relations. On the right, we show that semantic relations change by translation. **पय** translates to *liquid* and *water*, and their respective semantic relations with *water* differ. 71
5.2 The English path between *animals* and *pigs* has three edges: [`X/NOUN/nmod/>`, `species/NOUN/root/`$\wedge$, and `Y/NOUN/nmod/>`]. The path between animals and **सुअर** is defined as a combination of the English path and the Hindi path between **जानवरों** and **सुअर**. 74

5.3 Illustration of weakly supervised training: For a given English example (*cat*, *animal*), we generate predictions $\hat{\mathbf{l}}^{e \to e}$ using the monolingual English teacher model. Simultaneously, we also generate predictions $\hat{\mathbf{l}}^{e \to f}$ using the cross-lingual *student* model after translating one of the two English words using a dictionary. The cross-lingual classifier attends to all translation candidates and predicts a class based on a weighted average of their features. The loss is defined as CROSS-ENTROPY($\hat{\mathbf{l}}^{e \to e}$, $\mathbf{l}$) + CROSS-ENTROPY($\hat{\mathbf{l}}^{e \to f}$, $\mathbf{l}$) + KL-DIVERGENCE($\hat{\mathbf{l}}^{e \to f}$, $\hat{\mathbf{l}}^{e \to e}$). 75

7.1 Learning curves on the validation set for English-French models (mean of 3 runs/model). The SEMANTIC SIMILARITY model outperforms other models throughout training, including the one trained on all data. 117

7.2 Brevity penalties on the validation set for English-French models. . . 119

Chapter 1: Introduction

Resources such as parallel corpora and bilingual dictionaries serve as cornerstones of multilingual and cross-lingual natural language processing (NLP). These resources, which typically consist of words and sentences aligned with their translations in one or more languages, have a long history of use in building machine translation (MT) systems (including rule-based (Dugast et al., 2007), statistical (Brown et al., 1993; Yamada and Knight, 2001), and neural models (Bahdanau et al., 2014; Vaswani et al., 2017)). Beyond translation, they have also been used to facilitate cross-lingual learning by transferring labeled data (Hwa et al., 2005; Mayhew et al., 2017; Yarowsky et al., 2001) and trained models (Kozhevnikov and Titov, 2013; McDonald et al., 2011) from one language to another for many NLP tasks such as part-of-speech tagging, named entity recognition, semantic role labeling and syntactic parsing.

A common assumption in most of these works is that all translations recorded in these resources are *semantically equivalent*, *i.e.* the meaning of a word or a sentence is faithfully reproduced in its translation. However, this is often not the case. The meaning of a translation frequently *diverges* from that of the source text, and often does so in systematic ways. In this thesis, we focus on such mismatches

in meaning in multilingual settings where we expect the text to be aligned to each other, *i.e.* a text is paired with its corresponding translation. We refer to such mismatches as *cross-lingual semantic divergences*.

Definition 1. Cross-lingual Semantic Divergences: Mismatches in meaning in text and its translation which are expected to be aligned across languages and equivalent in meaning.

Cross-lingual semantic divergences arise naturally as a result of the translation process. Human translators apply a variety of strategies while translating a text from one language to another and many such strategies cause the meaning of the translation to diverge from the source text. For instance, Baker (2011) lists eight strategies used by professional translators to cope with issues that arise while translating, some of which are not meaning preserving *e.g.* replacing a word by a more general or a more specific word, or omitting parts of the source text to make the translation clear. As such, semantic divergences are found not only in automatically created noisy parallel corpora and dictionaries, but also in curated resources such as test corpora used to evaluate MT systems and dictionaries created by humans. Table 1.1 shows examples of divergent pairs selected from two corpora created in different ways—one by automatically aligning TV/movie subtitles from different languages (Tiedemann, 2007), and the other manually by professional translators (Callison-Burch et al., 2012). Divergences also occur in comparable corpora that describe the same topic in different languages, such as news or Wikipedia articles. For instance, the English Wikipedia article of actor David Tennant states

Divergent sentences in OpenSubtitles	
`en`	someone wanted to cook bratwurst.
`fr`	vous vouliez des saucisses grillées.
`gl`	you wanted some grilled sausages.
`en`	i don't know what i'm gonna do.
`fr`	j'en sais rien.
`gl`	i don't know.
`en`	- has the sake chilled? - no, it's fine.
`fr`	- c'est assez chaud?
`gl`	- it is hot enough?
`en`	you help me with zander and i helped you with joe.
`fr`	tu m'as aidée avec zander, je t'ai aidée avec joe.
`gl`	you helped me with zander, i helped you with joe.
Divergent sentences in newstest2012	
`en`	i know they did.
`fr`	je le sais.
`gl`	i know it.
`en`	the female employee suffered from shock.
`fr`	les victimes ont survécu leur peur.
`gl`	the victims have survived their fear.

Table 1.1: Parallel sentences exhibit semantic divergences, as can be seen in these examples (English sentence (`en`), French sentence (`fr`) and its gloss (`gl`)) observed in a random sample of OpenSubtitles and the newstest2012 test set from WMT evaluations.

"*David Tennant is a Scottish actor*", while the corresponding French article says "*David Tennant, est un acteur britannique*" ("David Tennant is a British actor"). Thus, the English article is more specific than the French article.

Automatically identifying cross-lingual semantic divergences provides semantic insight into multilingual text by precisely characterizing differences in meaning between words and sentences in different languages. Knowledge of divergences encoded in resources can help in better exploiting them for multilingual tasks that rely on such resource, such as (but not limited to) machine translations. Identifying

divergences in output of automatic methods provides a window into the working of such models, and an opportunity to further improve them.

1.1 Thesis Statement

Translation is not always meaning preserving which leads to cross-lingual semantic divergences that affect multilingual NLP tasks. Detecting such divergences requires ways of directly characterizing differences in meaning across languages through novel cross-lingual tasks, as well as models that account for translation ambiguity and do not rely on expensive, task-specific supervision.

1.2 Roadmap

We support the claims of the thesis statement in three ways. We quantify how often semantic divergences occur in commonly used resources, motivate tasks and build models to recognize and characterize differences in meaning across languages, and use these models to understand the impact of divergences on downstream tasks.

We start by surveying the body of literature relevant to the thesis in Chapter 2. After establishing our definition of divergences, we discuss how translation is not always meaning preserving. We also look at the various ways in which semantic divergences manifest in data. To contextualize our work on identifying meaning differences between across languages, we discuss relevant work in computational modeling of lexical semantics.

In Chapter 3, we establish that semantic divergences occur frequently in bilin-

gual resources that record translations. We analyze data from three commonly used datasets—two parallel corpora and a bilingual dictionary—containing translations of different granularities, and show that almost 20-40% of examples are judged to be divergent by human annotators. This highlights the need for precise ways of characterizing differences in meaning across languages.

Chapters 4 and 5 introduce two new tasks for precisely characterizing differences in cross-lingual word meaning by identifying the *semantic relation* between two words in different languages. Additionally, we also introduce models for these tasks that do not rely on task-specific cross-lingual training data and respect the ambiguous nature of translation. We start in Chapter 4 by motivating the task of identifying cross-lingual hypernymy, as hypernymy is commonly observed in automatically induced translation pairs (Peirsman and Padó, 2011). Our solution to this task, BiSparse-Dep, uses sparse, bilingual word embeddings combined with an unsupervised scoring function (Geffet and Dagan, 2005). BiSparse-Dep embeddings are learned jointly in the two languages using monolingual corpora, and a bilingual dictionary that aligns the two languages. Experiments show that BiSparse-Dep identifies cross-lingual hypernymy more accurately than methods based on translation and existing cross-lingual embeddings, when evaluated on a challenging new dataset covering four languages paired with English. BiSparse-Dep is also robust when exposed to various resource-scarce settings.

In Chapter 5, we then expand to simultaneously classifying between multiple lexical relations defined under the natural logic framework (MacCartney and Manning, 2009) which have been useful in describing relations between English

paraphrases (Pavlick et al., 2015), and in downstream natural language inference systems (MacCartney and Manning, 2007). We present BiLexNet, a weakly supervised neural classifier that learns to predict cross-lingual semantic relations. BiLexNet is trained on monolingual examples of semantic relations and a bilingual dictionary, using a novel approach based on knowledge distillation. Evaluated on a test bed of English-Hindi and English-Chinese word pairs labeled with the correct semantic relations, BiLexNet outperforms methods that more naively rely on bilingual embeddings or dictionaries for translation and cross-lingual transfer. Crucially, both BiSparse-Dep and BiLexNet rely on bilingual dictionaries for cross-lingual transfer, and in both cases, these dictionaries capture translation ambiguity by allowing multiple translations for each word.

Chapter 6 focuses on identifying semantic divergences in parallel sentences. We introduce an approach based on neural semantic similarity that separates divergent examples from semantically equivalent examples, and does so more accurately than models based on surface features and word alignments. Importantly, this model also does not require manually annotated task-specific data, and thus can be trained for any language pair and domain with a parallel corpus.

Chapter 7 takes an extrinsic view and studies semantic divergences in the context of downstream tasks. First, we model word-level divergences in the context of bilingual dictionary induction using BiLexNet, and show that even when the input word pair is close in meaning, BiLexNet helps separate divergent translations from equivalent translations. Second, we show that divergences in parallel sentences slow down training of neural machine translation models and filtering these out

using the method introduced in the previous chapter makes training twice as fast with no loss in translation quality.

Finally, we conclude in Chapter 8 by summarizing our contributions, highlighting the shortcomings of the work described in this thesis, and suggesting future directions.

1.3 Contributions

The main contributions of this thesis can be divided into four areas:

- **Conceptual**: Recognizing that translation is not always meaning preserving, we introduce two novel cross-lingual tasks *viz.* cross-lingual hypernymy detection, and cross-lingual identification of lexico-semantic relations, with the aim of precisely characterizing differences in word meaning across languages. These tasks have been previously only studied in monolingual settings where the two words are in the same language or in transfer settings, where models from one language are ported to another language. Instead, we propose cross-lingual tasks with the objective of directly comparing words in two languages.

- **Models**: We design three methods to identify differences in meanings between words and sentences in two languages. The lack of labeled data discourages traditionally supervised methods, hence we do not rely on task-specific cross-lingual training. Importantly, these methods incorporate translation ambiguity as part of the modeling process. Specifically, to identify how word meanings are related, we introduce an unsupervised algorithm for cross-lingual hyper-

nymy detection, and a weakly supervised method for identifying cross-lingual semantic relations. Both these methods rely on monolingual and bilingual corpora, and bridge the gap between the two languages using a bilingual dictionary that preserves ambiguity by recording multiple translations for each word. Our third method is a deep neural model for identifying semantic divergences in parallel sentences that uses synthetic supervision where the negative examples are based on mismatches from parallel segments.

- **Data**: To evaluate our (and future) models, we collect and release several datasets. First, we provide a dataset for evaluating models of cross-lingual hypernymy which covers four language pairs (English-French, English-Arabic, English-Chinese, and English-Russian), and contains approximately 3000 positive examples of cross-lingual hypernyms, paired with an equal number of negative examples. Second, we introduce MultiLexRel, a dataset of almost 2000 English-Hindi and English-Chinese word pairs labeled with natural logic relations. Finally, we release a dataset of parallel sentence pairs labeled with binary divergence annotations. Our models and datasets are all available publicly at `https://github.com/yogarshi`.

- **Applications**: We evaluate the impact of semantic divergences and semantic relations between words on two downstream tasks and show that divergences affect the quality of automatically constructed dictionaries, and slow down neural MT training. We also show that we can filter out semantic divergences to improve training times of neural MT models.

Chapter 2: Background

2.1 Translation

In Chapter 1, we defined cross-lingual semantic divergences as mismatches in meaning in text and its translation, that is typically expected to be aligned across languages and equivalent in meaning. We also claimed that such divergences arise because translation is not always meaning preserving. Here, we support this claim by looking at different ways in which translation equivalence has been defined and how such definitions do not necessarily ensure the preservation of meaning between the source and target texts.

2.1.1 Meaning and Semantic Equivalence

In order to study semantic divergences, we must first establish what it means for two texts to have the same meaning or to be *semantically equivalent*. This requires laying down an operational definition of meaning.

Defining the meaning of a word has been a fundamental problem in the philosophy of language and lexical semantics (see Gasparri and Marconi (2015) and Murphy (2010, Chap. 2) respectively, for a comprehensive discussion). For the pur-

pose of this thesis, we are interested in a *denotative* definition of meaning, *i.e.*, we want a definition that that tells us what a word can and cannot refer to. Denotative meaning stands in contrast to *connotative* meaning, which refers to the looser semantic associations of a word, which do not form the core denotative meaning of a word.

We follow Murphy (2010, p. 36) and define the meaning of a word as its *word senses*. Further, again following Murphy (2010, p. 39), we define a *word sense* as an abstract representation that connects the word form to the concept as it exists in the world. This definition of word meaning is in contrast to a referential theory of word meaning, which defines the meaning of a word as the set of things that it can refer to (or "point out") in the world.

The sense-based definition of word meaning leads to a simple definition of lexical equivalence that is also language independent—two words are semantically equivalent iff they share at least one word sense. As a proxy for this abstract representation, we computationally represent words using distributional or vector space representations which are empirically estimated using contextual co-occurrences from a corpus. Such representations are based on the distributional hypothesis, which claim that the meaning of a word is defined on the basis of the words that it co-occurs with (Firth, 1957; Harris, 1954).

We define the meaning of larger units of texts (such as phrases and sentences) following the principle of compositionality, *i.e.* the meaning of a sentence, depends in some way on the words that constitute the sentence. Again, as a proxy for this compositional representation, we will use an empirical model of representing the

meaning of a sentence that uses recurrent neural networks to encode the meaning of a sentence based on its words. We call two sentences semantically equivalent iff they are paraphrases of each other *i.e.* they represent the same event or fact without additional explanation of the context in which these two sentences occur.

The definitions of semantic equivalence between words and sentences laid down here are practical notions that can be easily applied across languages. They are also well suited for not just skilled translators, but also to bilingual speakers of both the source and target language such as the non-expert bilinguals who provide annotations in this work (Chapter 3 as well as Sections 4.4 and 5.3).

2.1.2 Translation Equivalence is not Semantic Equivalence

Defining *translation equivalence* has been a fundamental (and often controversial) problem in the translation studies literature. Broadly, translation equivalence is used to indicate that the text being translated (or the *source language* text) and its translation (or the *target langauge* text) have some sort of "sameness" or "similarity" (Panou, 2013). Different ways of characterizing this similarity have given rise to various definitions of translation equivalence over the years.

Vinay and Darbelnet (1958) claim that equivalence is situational. They present equivalence as a procedure in which a situation described in the source text, is presented in the target language using different words.

In a seminal work, Nida (1964) posits the idea of *dynamic equivalence*. Under this definition, a translation aims to evoke the same response in readers of the

translation, as it did in readers of the source language text. He contrasts this with *formal equivalence*, where the focus is on translating more literally and maintaining the lexical choices and the grammatical structure of the source language text. This definition moves beyond faithfulness to the source text, and is the first to take into account the receptor of the target text into account. Newmark (1981) also supports such a view of equivalence, using the terms *semantic* and *communicative* equivalence respectively. Thus, such definitions of equivalence move beyond the semantics of the source and the target text and focus on the effects of the two texts on readers.

Similarly, House (1997) claims that a source text and a target text are equivalent only if they serve the same function. This definition adopts a more pragmatic view of equivalence and requires taking into account both textual and non-textual aspects of the source and the target text.

Finally, it has also been argued that a single definition of equivalence is impossible to define. Koller (1979) and Baker (2011) both define a taxonomy of equivalences and claim that different situations require preserving different kinds of equivalences, and it is the job of the translator to identify such needs.

These definitions make it clear that good translations do not necessarily convey the exact meaning as the source text *i.e.* translation equivalence is not necessarily semantic equivalence, as defined in the preceding subsection (Section 2.1.1).

2.1.3 Automatically Identifying Translation Correspondences using Alignment

Computationally, translation correspondences are identified using the process of *alignment*. Given a source text and its translation in the target language, alignment refers to the procedure of identifying which part of the source text is translated to which part of a target text (Wu, 2010). Alignment is a key component of the machine translation pipeline and pieces of texts that are aligned to each other are expected to be semantically equivalent.

The primary resource that machine translation systems rely on are *parallel corpora*. Parallel corpora contain a collection of original texts in a source language and their translations into a set of target languages. The translated documents are automatically segmented into sentences which are further automatically aligned to their corresponding source sentences, based on information such as sentence length and order in the documents (Brown et al., 1991; Gale and Church, 1991). The resulting sentence pairs form training examples for machine translation (MT) systems. While neural MT architectures directly learn translation models from these sentence pairs, statistical MT systems rely on word level translation lexicons automatically induced by aligning words within sentence pairs (Brown et al., 1993).

A common assumption in multilingual NLP is that translations identified using alignments are semantically equivalent. For example, MT systems are regularly trained on all aligned sentences from commonly used parallel corpora (Bahdanau et al., 2014; Vaswani et al., 2017, *inter alia)*, cross-lingual embeddings are trained

using word-aligned and sentence-aligned corpora with the assumption that aligned words and sentences are equivalent in meaning (Klementiev et al., 2012; Luong et al., 2015), and word alignments are commonly used to build translation lexicons (Koehn et al., 2007). In studying cross-lingual divergences, we focus on situations where this assumption of alignment implying semantic equivalence does not hold.

2.1.4 Semantic Divergences in Noisy Parallel Data

While perfectly valid translations can cause semantic divergences, such divergences also manifest in real-world data due to different kinds of noise. A large body of prior MT work has focused on detecting noise in parallel data, and identifying the impact of such noise on phrase-based and neural MT systems.

Goutte et al. (2012) show that phrase-based systems MT are remarkably robust to noise in parallel segments. When introducing noise by permuting the target side of parallel pairs, as many as 30% of training examples had to be permuted to degrade translation quality significantly (measured using the BLEU score (Papineni et al., 2002)). While such artificial noise does not necessarily capture naturally occurring divergences, there is evidence that data cleaning to remove real noise can benefit MT, including in low-resource settings (Matthews et al., 2014).

Advances in machine translation afforded by neural models are accompanied by newer and perhaps more increased concerns about noise in data and its impact on models. While the idea of crawling the web for automatically obtaining parallel data is not new (Resnik, 1999), there has been an increasing amount of focus on

building such corpora to satisfy data-hungry neural models (Koehn et al., 2018a; Schwenk et al., 2019; Tiedemann, 2007). To mitigate the impact of noise from such corpora, Koehn et al. (2018b) propose a shared task where participants develop methods to filter a large, noisy parallel corpus to a smaller sized corpus of high quality sentence pairs. Models trained on this smaller corpus are evaluated on a variety of test sets to measure their generalization capability across domains and genres. The more successful participants rely on large neural MT and language models as features for learning scoring functions for sentence pairs, highlighting the need for more lightweight approaches.

On the modeling side, neural MT models appear to be more sensitive to the nature of training examples than phrase-based models. Chen et al. (2016) suggest that neural MT systems are sensitive to sentence pair permutations in domain adaptation settings. Belinkov and Bisk (2017) demonstrate the brittleness of character-level neural MT when exposed to synthetic noise (random permutations of words and characters) as well as errors that can be made naturally by humans. Hassan et al. (2018) claim that even small amounts of noise has adverse effects on neural MT models, as they tend to assign high probabilities to rare events. The higher sensitivity of neural models makes it imperative to test whether and to what extent are such systems affected by semantic divergences. We return to this problem in Chapter 7.

2.1.5 Translation Divergences

Previous work in MT has focused on *translation divergences*, which have been defined as structural or syntactical differences between sentences that convey the same meaning (Barnett et al., 1991; Dorr, 1990). The key difference between semantic divergences and translation divergences is that the former refers to differences in meaning, while the latter reflects the fact that languages can encode the same meaning in different ways.

The study of translation divergences was pioneered by Dorr (1994) who describes seven types of translation divergences based on English, German, and Spanish along with an interlingua-based MT approach to handle such divergences. Later work has used this classification to study translation divergences between more distant language pairs such as English-Hindi (Dave et al., 2001) and Urdu-English (Saboor and Khan, 2010). However, applying a taxonomy defined using a set of European languages to more distant languages is restrictive since divergence phenomena differ across languages from different families. Moreover, defining an exhaustive set of divergences manually for each language pair is difficult and needs bilingual experts. To overcome these issues, Deng and Xue (2017) propose a data-driven approach for studying translation divergences. They utilize the large amounts of parallel data available to modern MT systems to semi-automatically identify and categorize translation divergences between English and Chinese. Their strategy consists of manually aligning parse trees in the two languages, and then using the

alignment between the non-terminal nodes of the parse trees to extract divergences.[1] Using such a data-driven approach reveals that translation divergences occurring in large-scale parallel data are far more diverse than the few categories described in previous work. They also show that divergences are captured by rules encoded in a syntax-based SMT framework (Chiang, 2007).

As it stands, there has been little work that investigates the impact of translation divergences in neural MT models. While neural models are known to produce more fluent output (Bentivogli et al., 2016; Koehn and Knowles, 2017), it remains to be seen whether and to what extent can they handle such divergences.

2.2 Divergences in Sentences Beyond Parallel Data

The work discussed so far looks at semantic divergences in the context of parallel corpora and machine translation. However, a larger body of work has looked at the broader notion of semantic equivalence between sentences and phrases, both within and across languages, with different ways of characterizing non-equivalence.

The task of *paraphrase identification* aims to identify whether two texts (sentences or phrases) have the same meaning or not (Dolan et al., 2004). A key resource in the context of this task is the Paraphrase Database (PPDB) which consists of millions of automatically extracted paraphrases in different languages (Ganitkevitch et al., 2013). This dataset consists of lexical, phrasal, and syntactic paraphrases derived from large bilingual parallel corpora using the pivoting technique proposed

[1]Manually instead of automatically to isolate real divergences caused by different syntactic realizations from artificial ones produced by erroneous parses.

by Bannard and Callison-Burch (2005). The intuition behind the technique is that two strings in the source language that translate to the same string in a target language can be assumed to be paraphrases of each other. However, PPDB is rife with divergences—Pavlick et al. (2015) show that the diversity of semantic relations found in word-aligned parallel corpora causes the pivoting technique to yield semantically divergent paraphrases. They (automatically) annotate the English PPDB with semantic relations defined under the natural logic framework (MacCartney and Manning, 2009). This labeling reveals that the largest English PPDB collection (PPDB-XXXL) which consists of 77.4M paraphrase pairs, contains less than 10% pairs which are truly equivalent. This does not diminish the utility of the other 90% pairs, but knowing the precise relation between these divergent pairs can help in using them more effectively for downstream tasks. The idea that not all paraphrases have to be logically equivalent has also been supported through linguistic definitions of paraphrases, which allow for *quasi-synonymity* or only approximate equivalence between paraphrases (Beaugrande and Dressler, 1981; Bhagat and Hovy, 2013).

Two texts that are not exactly equivalent or paraphrases of each other exhibit varying degrees of similarity. The task of identifying the *semantic textual similarity* softens the binary notion of equivalence assumed by paraphrase tasks. The objective instead is to assign a real valued score to a sentence pair that captures a graded similarity between the two sentences (Agirre et al., 2014; Corley and Mihalcea, 2005). However, STS only tells us to what extent two sentences differ, but not *how* they differ. One framework that explicitly characterizes the nature of semantic divergences is *textual entailment* (TE) (Dagan and Glickman, 2004). To recognize

textual entailment is to identify whether the meaning of one sentence (the premise or the text) implies the meaning of another sentence (the hypothesis). While motivated from logical entailment, the definition of textual entailment is more relaxed: a sentence t entails h if, typically, a human reading t would infer that h is most likely true. By virtue of this definition, textual entailment also eschews the symmetry inherent to paraphrase detection and STS, since entailment relations are asymmetric. In the following example,

– **Premise**: Raj bought a novel yesterday

– **Hypothesis**: Raj purchased a book

the premise entails the hypothesis, but the reverse is not true.

These various frameworks have also been used to study differences in meaning across languages. In multilingual settings, work on cross-lingual semantic textual similarity (Agirre et al., 2014) and cross-lingual textual entailment (Mehdad et al., 2010; Negri et al., 2012, 2013) characterizes semantic relations between sentences in different languages beyond translation equivalence. These tasks have similar goals as their monolingual counterparts, with the key difference being that the input consists of a pair of words, phrases or sentences in different languages. Recent work has also promoted work on cross-lingual *transfer* for such tasks, where models trained on English are evaluated on other languages (Agić and Schluter, 2018; Conneau et al., 2018). Cross-lingual models share core intuitions, relying either on MT to transfer the cross-lingual task into its monolingual equivalent (Jimenez et al., 2014; Zhao et al., 2013), or on features derived from MT components such as translation

dictionaries and word alignments (Lo et al., 2016; Turchi and Negri, 2013).

Models based on deep neural networks have also been proposed for monolingual and cross-lingual versions of the tasks (He et al., 2015; Rocktäschel et al., 2015; Tai et al., 2015, *inter alia*). These models are generally more accurate than their non-neural counterparts (He and Lin, 2016). Naturally, improved performance comes with reliance on large amounts of training data. This restricts the direct application of neural models in cross-lingual settings, where labeled training data is scarce to non-existent. We return to neural models and the question of their training in Chapter 6 where we discuss a deep neural architecture for semantic similarity that we re-purpose for detecting divergences in parallel data, and show an effective way of training such a model without human annotated cross-lingual data.

A key limitation of datasets proposed for cross-lingual STS (Cer et al., 2017) and TE (Conneau et al., 2018; Negri et al., 2012, 2013) is that they are prepared by translating one side of existing datasets into the language of interest, specifically for the purpose of the task. Such techniques rely on the assumption that translation is meaning preserving and that semantic relations between words and sentences are preserved across languages. As discussed in Section 2.1.2, this assumption is not always valid, and in this thesis we build datasets of divergences that consist of naturally occurring cross-lingual pairs drawn from a parallel corpus (Chapter 3).

2.3 Lexical Divergences and Semantic Relations

In the previous two sections, we saw various ways in which divergences arise in translations, and we saw how various frameworks have been introduced to characterize differences in meaning between sentences. In this section, we look at semantic divergences between words.

2.3.1 Non-equivalence of Lexical Translations

The observation that a word and its translation often do not cover the exact same semantic space has been recognized and exploited in various ways. Even withing a single language, it has been widely argued that true synonymy is rare, if it exists at all. The principle of contrast by Clark (1987) represents one extreme end of such argument as it claims that if two words differ in form, they differ in meaning. On the other hand, Church et al. (1994) argue that many (but not all) synonyms can be identified by their ability to substitute for each other in broader context. Edmonds and Hirst (2002) claim that true meaning equivalence (or true synonymy) between words is very rare. They instead argue that synonyms are more likely to be *near-synonyms*. Near-synonyms are very similar, but not identical, in meaning. They are not fully inter-substitutable, but instead vary in their shades of denotation, connotation, implicature, emphasis, or register.

Edmonds and Hirst (2002) also claim that near-synonymy is the norm for lexical choice in translation, *i.e.* the word in the target language that is closest to that in the source text is more often a *near-synonym* rather than an exact synonym.

Words are also often translated non-literally due to lexical gaps or decisions made by translator (Bentivogli and Pianta, 2000; Santos, 1990). For instance, Baker (2011) suggests that a common strategy used by translators is to translate a word to a more general word, if an exact translation for the original word cannot be found. Consider the example below:

- **Source Text** : Shampoo the hair and lightly towel dry.
- **Target Text** : Lavar el cabello y frotar ligeramente con una toalla.
- **Gloss** : *Wash hair and rub lightly with a towel.*

Here, the translator has chosen to replace the word "shampoo" by the more general word "wash", which leads to a semantic divergence (you can *wash* lots of things, but you can only *shampoo* hair). This shows that semantic divergences often arise as a result of differences in meaning between words and their translations, and often such words and their translations are related in systematic ways.

A related observation is made by Peirsman and Padó (2011) who claim that that many pairs in automatically generated translation lexicons exhibit semantic relations beyond synonymy. In an analysis of a German-English lexicon, they observe that only about half of a random sample of pairs exhibit synonymy, and about 20% evoke well-defined taxonomic relations other than synonymy *i.e.* relations that are covered by a taxonomy or ontology such as WordNet (Fellbaum, 2010; Miller, 1995). These relations include:

1. **Hypernymy**: Is-A relation; *e.g.* Dramatiker ('playright') - writer (writer is a hypernym of Dramatiker)

2. **Hyponymy**: Reverse of hypernymy; *e.g.* Kunstwerk ('work of art') - painting (painting is a hyponym of Kunstwerk)

3. **Co-hyponymy**: Words with a common hypernym; *e.g.* Straßenbahn ('tram') - bus

4. **Antonymy**: Opposites; *e.g.* Inneres ('interior') - exterior

Inspired by this observation, they call for work on identifying such relations in cross-lingual settings, but there has been little work on such tasks. Our work (Chapters 4 and 5) aims to fill this gap by defining cross-lingual tasks that directly characterize differences in meaning between words in two languages by identifying the semantic relation between them. We also introduce datasets for this task consisting of bilingual words pairs that are directly annotated for the semantic relations of interest, and are not simply translations of monolingual datasets as in prior work (Glavaš and Vulić, 2018).

While work in cross-lingual settings has been limited to multilingual taxonomies (Bond and Foster, 2013; Navigli and Ponzetto, 2012), semantic relations beyond synonymy have been very well-studied in monolingual settings. Our cross-lingual models in Chapters 4 and 5 share several core ideas with methods for the monolingual task, so we briefly discuss the various relations and methods for identifying them within a single language.

2.3.2 Lexico-semantic Relations

The term *semantic relations* is used to denote meaningful, well-defined associations between two or more concepts, entities or sets of entities. When these relations are studied on the basis of words used to represent the concepts, then they are called lexical relations, or *lexico-semantic relations*. In this thesis, we use these two terms interchangeably, since we will always focus on relations between concepts represented by words.

Semantic relations have been studied by Cruse (1986) under the umbrella of congruence relations between lexical items. He defines four congruence relations using elementary relations from set theory. The four congruence relations defined over classes A and B (with their corresponding semantic relations) are:

1. **Identity:** A and B have the same members
2. **Inclusion:** B is wholly included in A
3. **Overlap:** A and B have members in common but each has members not found in the other
4. **Disjunction**: A and B have no members in common

The congruence relations defined above are further used by Cruse (1986) to a fundamental set of lexical relations, which have also been studied computationally. Inclusion, as defined by Cruse (1986), corresponds to the **hypernymy/hyponymy** relations. A word w_1 is a hypernym of another word w_2, if the concept represented

by w_2 can be claimed to be a kind of w_1 *e.g.* *author* is a hypernym of *Shakspeare*, and *Shakespeare* is a hyponym of *author*. A closely associated relation to hypernymy is **co-hyponymy**. Two words are said to be co-hyponyms if they share a common hypernym *e.g.* *Shakespeare* and *Austen* are co-hyponyms because they both share a common hypernym *viz.* *author*. **Antonymy** captures opposites *e.g.* *hot* and *cold*. Both antonymy and co-hyponymy can be thought of as sharing the disjunction congruence relation. Finally, **meronymy** (and its reverse **holonymy**) captures part-whole relationships between concepts : *engine* is a meronym of *car*, *finger* is a meronym of a *hand*. While these relations capture principled associations between words, they only exist between a small set of words. Boyd-Graber et al. (2006) propose identifying the **evocation** relation between concepts *i.e.* to what extent does one concept bring to mind another. Unlike the other relations, evocation is a weighted relation since different concepts can have different strengths of evocations with respect to a specific concept. Recent work by Vulić et al. (2017) also adds a weight to the hypernymy relation, with the claim that not all hypernym relations are equally significant.

An overarching concept that attempts to unify several of these relations is that of **lexical entailment** which was introduced by Geffet and Dagan (2004) to capture the notion of meaning-preserving substitutability. Informally, a word w lexically entails another word v, if w can substitute for v in some contexts, while implying v's original meaning. Lexical entailment generalizes over the taxonomic relations discussed above—synonyms typically entail each other, hyponyms entail their hypernyms, while entailment holds for meronymy in only certain cases. Resnik

Natural Logic Relation	**Taxonomic Relation**	**PPDB Label**	**Monolingual Example**
Equivalence	Synonymy	Equivalence	dog, canine
Forward Entailment	Hypernymy	Forward Entailment	crow, bird
Reverse Entailment	Hyponymy	Reverse Entailment	bird, crow
Negation	Antonymy	Exclusion	good, evil
Alternation	Co-hyponymy	Exclusion	dog, cat
Independence	Other	Other	hungry, hippo

Table 2.1: A unified summary of taxonomic semantic relations, natural logic relations (MacCartney and Manning, 2007), and the re-definition of the natural logic relations for PPDB (Pavlick et al., 2015), along with examples.

(1993) advances a similar notion of *plausible entailment* to define synonymy: two words share a meaning if there is a representative context in which they are mutually substitutable without changing the inferences that one can draw about that context.

Variations on some of these relations have been defined under the natural logic framework defined by MacCartney and Manning (2007, 2009) to perform textual inference directly over natural language without using formal logic representations. The natural logic relations were further refined by Pavlick et al. (2015) to identify relations between paraphrases in PPDB. The negation and alternation relations as defined by MacCartney and Manning (2009) are replaced by the broader notion of exclusion, while the cover relation is entirely dropped as its practical utility is unclear. We use this re-definition in our study of cross-lingual semantic relations (Section 6.2). These relations and their correspondence with taxonomic relations are summarized in Table 2.1.

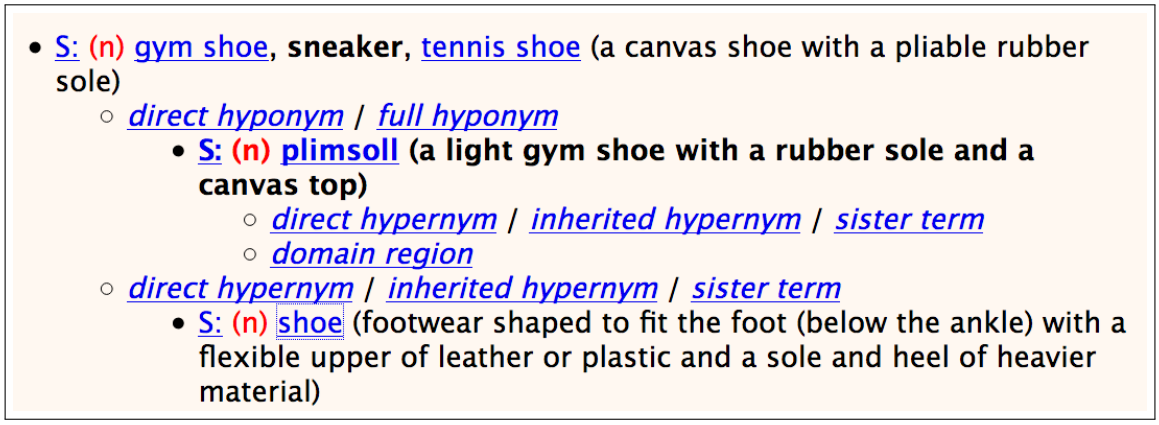

Figure 2.1: Example of a WordNet synset (*gym shoe*, *sneaker*, *tennis shoe*), along with its hypernym (*shoe*), and its hyponym (*plimsoll*) synsets

2.3.3 Automatic Methods for Detecting Hypernymy and Hyponymy

The most well-studied relations from those discussed in the previous section are the dual hypernymy-hyponymy relations. Fundamentally, these relations are defined between word meanings or concepts (Miller et al., 1990) : a concept lexicalized as L_1 is a hypernym of another concept lexicalized as L_0 (or L_0 is a hyponym of L_1), if L_0 can be claimed to be a kind of L_1. One of the earliest contributions to the computational study of hypernyms and hyponyms was WordNet, a taxonomy that organizes words in English into unordered sets of synonyms that each expresses a distinct concept (Miller et al., 1990). These *synsets* are then linked using various semantic relations, of which hypernymy and hyponymy are the most frequent. Figure 2.1 shows an example synset with its hypernym and hyponym synsets.

Defining such relations simply on the basis of concepts organized in a taxonomy is restrictive, as this does not allow inferring relations between new words or words from a different domain that are not already present in the taxonomy. This has

encouraged development of automatic methods for identifying hypernymy between two word types. There are two main family of methods for identifying hypernymy:

Pattern-based The earliest approaches for automatically identifying hypernyms suggested that noun phrases connected by specific textual patterns are indicative of hypernymy (Caraballo, 1999; Hearst, 1992) *e.g.* "such *authors* as *Shakespeare*" indicates that *author* is a hypernym of *Shakespeare*. Snow et al. (2005) extend such approaches to take into account syntactic patterns extracted from dependency parses, which allow capturing of more complex long-distance phenomena. Such approaches have also been extended to languages other than English (Lefever et al., 2014; Yildirim and Yildiz, 2012). Path-based methods are limited by their recall, since they model the hypernymy relation based on the joint occurrence of the two input words, and thus require both words to occur together in a sentence in the given corpus.

Distributional Distributional methods, on the other hand, base their predictions on separate contexts for each of the two words. Each input pair is represented using a combination of the vector representations of the two words, and a prediction is made based on this input representation. Distributional methods do not suffer from the coverage issue that pattern-based approaches suffer from as they do not require the two input words to co-occur. However, they are less precise at identifying the hypernymy relation, and more successful at detecting broad semantic similarity.

Distributional methods can also be further divided into unsupervised and su-

pervised techniques. Unsupervised methods take in as input the vector space representations (or, word embeddings) of the two words as features and output a real valued score which indicates the degree of hypernymy. The scoring functions used by these approaches are motivated by specific linguistic hypotheses. One such idea underlying several scoring functions (Clarke, 2009; Kotlerman et al., 2009; Lenci and Benotto, 2012) is the *Distributional Inclusion Hypothesis* which posits that the prominent context features of a hyponym are expected to be included in those of its hypernym (Geffet and Dagan, 2005). Scoring functions based on this hypothesis have been highly successful at detecting hypernymy (Shwartz et al., 2017). The distributional inclusion hypothesis serves as a key modeling hypothesis for our BiSparse-Dep model for cross-lingual hypernymy identification (Section 4.2).

Supervised methods for hypernymy detection generally outperform unsupervised approaches, but they require labeled training data *i.e.* word pairs which are known to be hypernyms/hyponyms of each other. Early supervised approaches have relied on simple operations over the word embeddings of the two words as input features to a linear classifier or an SVM with a polynomial kernel (Fu et al., 2014; Roller et al., 2014; Weeds et al., 2014). Prompted by concerns that such methods only learn *prototypical* properties of individual words (such as the fact that *animal* is likely to be a hypernym), and not relations between the word pair (Levy et al., 2015), subsequent techniques have incorporated non-linear transformations (Glavaš and Ponzetto, 2017). More recent supervised methods attempt to combine the precision of pattern-based approaches with the coverage of embedding based methods (Roller and Erk, 2016; Shwartz et al., 2016). For instance, Shwartz et al. (2016) show that

by using an integrated neural network which encodes patterns between words using an LSTM, and combines this encoded representation with the distributional representations of the two words, they can more successfully identify hypernyms than purely pattern-based or distributional methods.

2.3.4 Automatic Methods for Other Relations

Automatic methods for identifying relations beyond hypernymy also use both pattern-based (Berland and Charniak, 1999; Chklovski and Pantel, 2004; Nguyen et al., 2017) and distributional approaches (Yih et al., 2012). However, most methods target a single relation and isolate instances of that relation from those of other relations. In general, methods that deal with multiple semantic relations are fewer (Pantel and Pennacchiotti, 2006; Pennacchiotti and Pantel, 2006; Turney, 2008), and recent shared tasks have shown that this is a challenging problem, especially when ontologies and other structured resources are not available, and models are trained only on raw corpora (Santus et al., 2016).

Shwartz and Dagan (2016b,c) generalize their integrated method for detecting hypernymy (Shwartz et al., 2016) and show that they can more successfully distinguish between multiple relations than other approaches. This model, called LexNet, is a starting for our model for identifying cross-lingual semantic relations (Chapter 5). However, LexNet is fully supervised and needs labeled data for training. This is an unrealistic assumption in cross-lingual settings, and we discuss how we train a comparable cross-lingual model without direct supervision (Section 5.2).

2.3.5 Semantic Relations across Languages

Work on lexico-semantic relations discussed above has largely focused on a single language, which is more often than not English. In recent years, however, there has been some investigation on cross-lingual *transfer* of models for semantic relations. This line of works asks whether we can use models trained in a high resource language with labeled data (say English) to identify semantic relations in other languages (Glavaš and Vulić, 2018; Roth and Upadhyay, 2019). The assumption of availability of only high-resource training data is shared by our model (Section 5.2). However, we make predictions between words in two different languages, while the aforementioned works still focus on a single language.

The development of linked multilingual resources such as Babelnet (Navigli and Ponzetto, 2012) and the Open Multilingual WordNet (Bond and Foster, 2013) also provides a way to identify relations across languages, but just as in monolingual WordNet, these resources are limited by domain and vocabulary, and expensive to create and maintain. In this thesis, we generalize traditionally monolingual tasks to cross-lingual settings by providing datasets and benchmarks for cross-lingual hypernymy (Chapter 4) and cross-lingual semantic relations (Chapter 5). We build models that can automatically identify such relations without relying on expensive cross-lingual ontologies or labeled training data.

2.3.6 Utility in downstream tasks

Apart from identifying intrinsic relations between words, lexico-semantic relations have served a variety of downstream NLP tasks. Knowledge of hypernyms has proven to be useful in many different NLP tasks, such as textual entailment Dagan et al. (2013), coreference resolution (Ponzetto and Strube, 2006), relation extraction (Demeester et al., 2016), and question answering (Huang et al., 2008). More broadly, knowing whether and how two words related is useful for automatic generation of thesauri (Grefenstette, 1994), building domain specific ontologies (Zouaq and Nkambou, 2008) and generating paraphrases (Madnani and Dorr, 2010).

The ability to detect semantic relations across languages can also serve as a building block in corresponding cross-lingual tasks, including cross-lingual textual entailment (Negri et al., 2012, 2013). It can also help in constructing multilingual taxonomies (Fu et al., 2014) by helping organize lexicons across multiple languages, and in evaluating Machine Translation output (Pado et al., 2009) by allowing direct comparison of a translated text with the source. As a case study (Chapter 7), we demonstrate that our model for recognizing semantic relations can help in filtering semantically equivalent cross-lingual pairs from divergent pairs, and improve a bilingual dictionary model based on word embeddings.

2.4 Summary

This chapter supports the central claim of this thesis, *viz.* translation is not always meaning preserving, by presenting definitions of translation equivalence from

prior work in translation studies, and establishing how these definitions do not guarantee semantic equivalence. We also look at other ways in which divergences manifest in data (such as through noise), and various frameworks which have been proposed to characterize differences in meaning across and within languages, both between words and sentences.

Chapter 3: Quantifying and Analyzing Divergences in Bilingual Resources

We start our study of cross-lingual semantic divergences by focusing on a practical question: how frequently do semantic divergences occur in resources that record translations? We answer this question by annotating divergences in translations drawn from a bilingual dictionary and two sentence-aligned parallel corpora. These resources contain translations of different granularities, and in both cases 20–40% of examples studied are found to be semantically divergent. Examples reveal that these divergent translations are diversely related to the source words or sentences.

3.1 Divergences in Bilingual Dictionaries

Data Selection We start by analyzing divergences in high-quality bilingual dictionaries created by annotators on Mechanical Turk and spanning 100 languages (Pavlick et al., 2014).[1] Each dictionary contains ∼10000 words for a particular language, with multiple English translations for each word. Strict quality control has been followed in the creation process—a subset of the annotations have been

[1] `https://www.mturk.com`

compared against gold standard translations, copy-pasting from automatic services (*e.g.* Google Translate) has been checked for and filtered out, and translation quality has been measured and only translations that meet a quality threshold have been retained. These dictionaries have been used as a gold standard test bed to evaluate automatic methods for inducing bilingual dictionaries (Irvine and Callison-Burch, 2017), making it all the more important to understand the nature of translations captured by them.

Annotation Protocol We choose a random subset of 500 Hindi words and their translations from the dictionary to study the presence of divergences. We only choose lexical *i.e.* single word translations. These words pairs are annotated by three annotators, who are native or near-native speakers of Hindi, and fluent speakers of English. Annotators perform a binary decision task of identifying whether the translation pair is equivalent in meaning or not. Specifically, they are asked whether the meaning of the English word and the Hindi word is the same.

Annotation Analysis Figure 3.1 shows the breakdown of results of the annotation with respect to annotator agreement. Annotators largely tend to agree—for 75% examples all three annotators assign the same label. 63% examples (314/500) are labeled as being equivalent by all three annotators. This is not surprising given that this dictionary has been carefully created to ensure high quality translations. Factoring in examples that are annotated as being equivalent by two out of three annotators, this number goes up to almost 80% (393/500). However, this implies that

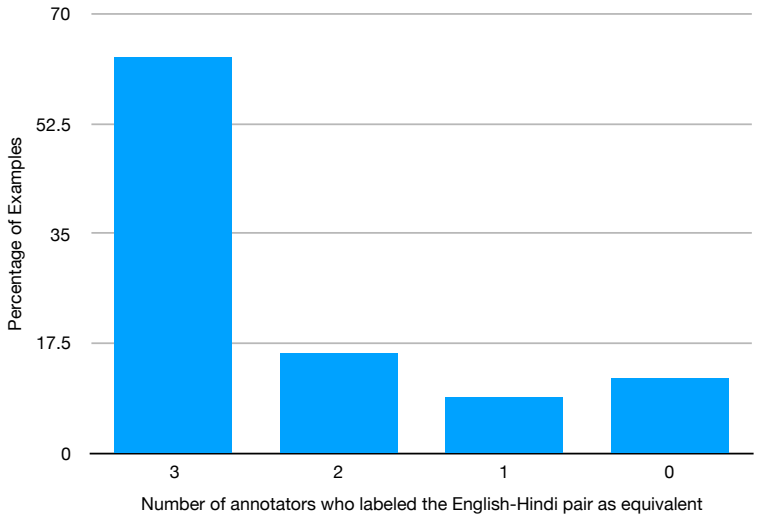

Figure 3.1: Breakdown of a random sample of 500 examples from a Hindi-English bilingual dictionary with respect to annotator agreement.

for about 20% examples (107/500), two annotators agreed on the example being not equivalent.

Table 3.1 shows some random examples of Hindi words and their English translations that are labeled as being not equivalent by two out of three annotators. These divergent English translations are related to the Hindi words in different ways. Several are related by taxonomic relations such as antonymy (आयात, *export*), hypernymy/hyponymy (नन्दा, *surname*), and co-hyponymy (जैवमण्डल, *atmosphere*). Others are associated words (घूर्णन, *swivel*) or noisy translations (तेलुगू, *telgu*). There are some annotation errors as well (*e.g.* वाहिनी and *duct* mean the same but are annotated as divergent), but looking at the data reveals that these are rare.

Hindi Word	English Translation	Gloss of Hindi Word
आयात	export	import
पतले	soft	thin
नन्दा	surname	nanda (a specific surname)
तेलुगू	telgu	telugu
निडारिया	fishes	cnidaria
जैवमण्डल	atmosphere	biosphere
पं	west	pan
घूर्णन	swivel	rotation
वाहिनी	duct	duct / channel

Table 3.1: Examples of Hindi-English pairs labeled as divergent by annotators.

3.2 Divergences in Parallel Sentences

Data Selection Having looked at examples of divergences in bilingual dictionaries, we now turn our attention to parallel sentences. We crowdsource annotations of English-French sentence pairs to assess how frequent semantic divergences are in parallel corpora. We draw examples for annotation randomly from two English-French corpora, using a resource-rich and well-studied language pair, and for which bilingual annotators can easily be found. The **OpenSubtitles** corpus contains 33M sentence pairs based on translations of movie subtitles. The sentence pairs are expected to not be completely parallel given the many constraints imposed on translations that should fit on a screen and be synchronized with a movie (Lison and Tiedemann, 2016; Tiedemann, 2007), and the use of more informal registers which might require frequent non-literal translations of figurative language. The **Common Crawl** corpus contains sentence-aligned parallel documents automatically mined from 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 resulting corpus of 3M sentence pairs is noisy, yet extremely useful to improve translation quality for multiple language pairs and domains (Smith et al., 2013).

Annotation Protocol Divergence annotations are obtained via Crowdflower.[2] Since this task requires good command of both French and English, we rely on a combination of strategies to obtain good quality annotations, including Crowdflower's internal worker proficiency ratings, geo-restriction, reference annotations by a bilingual speaker in our lab, and instructions that alternate between the two languages (Agirre et al., 2016).

Annotators are shown an English-French sentence pair, and asked whether they agree or disagree with the statement "the French and English text convey the same information." We do not use the term "divergent", and instead frame the question in accordance with the definition of semantic equivalence set forth in Chapter 2. We set up two distinct annotation tasks, one for each corpus, so that workers only see examples sampled from the same corpus in a given job. Each example is shown to five distinct annotators.

Annotation Analysis Forcing an assignment of divergent or equivalent labels by majority vote yields 43.6% divergent examples in OpenSubtitles, and 38.4% in Common Crawl. Fleiss' Kappa indicates moderate agreement between annotators (0.41 for OpenSubtitles and 0.49 for Common Crawl). This suggests that the annotation protocol can be improved, perhaps by using graded judgments as in Semantic

[2] `http://crowdflower.com`

Textual Similarity tasks (Agirre et al., 2016), or for sentence alignment confidence evaluation (Xu and Yvon, 2016).

Current annotations are nevertheless useful, and different degrees of agreement reveal nuances in the nature of divergences (Table 3.2). Examples labeled as divergent with high confidence (lowest block of the table) are either unrelated or one language misses significant information that is present in the other. Examples labeled divergent with lower confidence contain more subtle differences (*e.g.* "what does it mean" in English vs. "what are the advantages" in French).

3.3 Summary

This chapter established that semantic divergences occur frequently in bilingual resources that record translations. 20% of Hindi-English word pairs in a bilingual dictionary, and 40% of English-French parallel sentences from two different corpora were annotated as being semantically divergent by annotators. This exercise reveals that divergences occur in corpora that record translations of different granularities. Divergences discovered cover a wide spectrum, including subtle differences in meaning, well-defined taxonomic relations, as well as noisy translations.

Datasets annotated in this chapter will later be used to estimate the accuracy of divergence detection models (Chapter 6) and to gauge the utility of models for downstream tasks (Chapter 7). Before that, in the next three chapters, we turn to the problem of building models that can detect semantic divergences and differences in meaning across languages.

		Equivalent with High Agreement ($n = 5$)
subs	en	the epidemic took my wife, my stepson.
	fr	l'épidémie a touché ma femme, mon beau-fils.
	gl	the epidemic touched my wife, my stepson.
cc	en	to instantly check availability for all san sebastián hostels, use the form on the left of the page.
	fr	pour vérifier la disponibilité de toutes les auberges à saint-sébastien, utilisez le formulaire à gauche.
	gl	to verify the availability of all the hostels in san sebastián, use the form on the left.
		Equivalent with Low Agreement ($n = 3$)
subs	en	she was a kind person, then, was she?
	fr	c'était quelqu'un de gentil, non ?
	gl	it was someone nice, no?
cc	en	cancellation policy: if cancelled up to 28 days before date of arrival, no fee will be charged.
	fr	conditions d'annulation : en cas d'annulation jusqu'à 28 jours avant la date d'arrivée, l'hôtel ne prélève pas de frais sur la carte de crédit fournie.
	gl	cancellation conditions: in case of cancellation up to 28 days before arrival date, the hotel does not charge fees from the credit card given.
		Divergent with Low Agreement ($n = 3$)
subs	en	i tried to keep things nice and civil... but, hey, 25,000 for three suits?
	fr	je voulais que tout se passe gentiment, mais dites... 25 000 dollars pour 3 costumes ?
	gl	I wanted that everything goes nicely, but say... 25 000 dollars for 3 suits?
cc	en	what does it mean when food is "low in ash" or "low in magnesium"?
	fr	quels sont les avantages d'une nourriture "réduite en cendres" et "faible en magnésium" ?
	gl	what are the advantages of a food "low in ash" or "low in magnesium"?
		Divergent with High Agreement ($n = 5$)
subs	en	rabbit? if i told you it was a chicken, you wouldn't know the difference.
	fr	vous croirez manger du poulet.
	gl	you think eat chicken
cc	en	you need food to fuel your body to help you push further, to run faster, to perform at the highest possible level.
	fr	ce mois-ci, pourquoi ne pas vous fouetter le pâté aux patates parfait ?
	gl	this month, why not whisk yourself pâté potato perfect?

Table 3.2: Randomly selected sentence pairs (English (en), French (fr) and gloss of French (gl)) annotated as divergent or equivalent, with high and low degrees of agreement between the 5 annotators. Examples are taken from the OpenSubtitles (subs) and Common Crawl (cc) corpora.

Chapter 4: Identifying Cross-lingual Hypernymy using Sparse Bilingual Embeddings

The notion of cross-lingual semantic divergences between words is tightly coupled with the problem of identifying how words in different languages are related to each other. Knowing when two words differ in meaning is useful in identifying divergences, but the knowledge of *how* they differ can help in more precisely characterizing differences in word meaning across languages. Thus, in the next two chapters, we focus on identifying semantic relations between words across languages using algorithms that do not rely on task-specific cross-lingual training data. In the present chapter, we focus on identifying a single relation semantic relation, *viz.* hypernymy. We choose hypernymy as hypernyms/hyponyms are commonly found in automatically induced translation pairs (Peirsman and Padó, 2011). Besides, the lack of an equivalent target language word for a given source word often leads translators to choose a hypernym or a hyponym while translating (Baker, 2011; Chesterman, 1997). Hypernyms have also received significant attention in monolingual (mostly English) settings as a representation-agnostic way of modeling lexical semantics (Hearst, 1992; Lenci and Benotto, 2012; Shwartz et al., 2016; Snow et al., 2005; Weeds and Weir, 2003, *inter alia*). As such, there is a wealth of work on

modeling hypernymy which motivates approaches for the cross-lingual task.

Our emphasis in the next two chapters is *intrinsic*, *i.e.* we focus on building models for identifying semantic relations between words and evaluating how well various approaches perform at this task. We will return to the impact of such models vis-á-vis semantic divergences when we investigate the value of such models for downstream tasks in Chapter 7.

4.1 Cross-lingual Hypernymy : Challenges and Contributions

Building models that can robustly identify hypernymy across the spectrum of human languages is a challenging problem, that is further compounded in low resource settings. At first glance, translating words to English and then identifying hypernyms in a monolingual setting may appear to be a sufficient solution. However, this approach is impaired by its inability to capture translation ambiguity. For instance, the English words *cook*, *leader* and *supervisor* can all be hypernyms of the French word *chef*, as the French word does not have a exact translation in English covering its possible usages. However, translating *chef* to *cook* and then determining hypernymy monolingually precludes identifying *leader* or *supervisor* as a hypernyms of *chef*. Similarly, language-specific usage patterns can also influence hypernymy decisions. For instance, the French word *chroniqueur* translates to *chronicler* in English, but is more frequently used in French to refer to journalists (making *journalist* its hypernym).[1]

This motivates approaches that *directly* detect hypernymy in the cross-lingual

[1]All examples are from our dataset described in Section 4.4.

setting by extending distributional methods for detecting monolingual hypernymy. Limited training resources in cross-lingual settings make unsupervised methods more desirable than supervised hypernymy detection approaches. However, monolingual distributional methods cannot be applied directly to the cross-lingual task, because the vector spaces of two languages need to be aligned using a cross-lingual resource (a bilingual dictionary, for instance). Finally, state-of-the-art distributional approaches (Roller and Erk, 2016; Shwartz et al., 2017) for detecting monolingual hypernymy require syntactic analysis (*e.g.* dependency parsing), which may not available for many languages, raising the question of whether syntactic transfer from related languages (Zeman and Resnik, 2008) is a useful substitute in such situations.

We address these challenges using BiSparse-Dep—a family of robust, unsupervised approaches for identifying cross-lingual hypernymy. BiSparse-Dep uses a sparse, bilingual word embedding model learned from a small bilingual dictionary and a variety of monolingual syntactic context extracted from a dependency parsed corpus. We extensively evaluate BiSparse-Dep on a new crowd-sourced cross-lingual dataset for hypernymy detection, with over 2900 hypernym pairs, spanning four languages from distinct families—French, Russian, Arabic and Chinese. Our evaluation shows that BiSparse-Dep is more accurate than similar models which use window based contexts, or weaker baselines that simply rely on translations. Crucially for cross-lingual settings, BiSparse-Dep also exhibits robust behavior along multiple dimensions. In the absence of a dependency treebank for a language, it learns embeddings using a parser trained on related languages. When exposed to less monolingual data, or a lower quality bilingual dictionary, BiSparse-Dep

degrades only marginally. In all these cases, it compares favorably with models that have been supplied with all necessary resources, showing promise for low-resource settings. Our crowdsourced datasets are also publicly available for future work.

4.2 Unsupervised Identification of Cross-lingual Hypernymy

As in the monolingual case, we perform unsupervised identification of cross-lingual hypernymy using a scoring function which quantifies the *directional* similarity of an input word pair. A variety of functions have been introduced to quantify the directional relationship between two words, given feature representations of the two words (Lenci and Benotto, 2012; Lin, 1998; Weeds and Weir, 2003). A key idea underlying several functions is the **Distributional Inclusion Hypothesis**: given feature representations of the contexts of two words u and v, v is a hypernym of u if all features of u tend to appear within the features of v (Geffet and Dagan, 2005). Scorers based on the distributional inclusion hypothesis have been found to accurately distinguish hypernymy from other relations (Shwartz et al., 2017).

Specifically, we use *BalAPinc* to score word pairs for hypernymy (Kotlerman et al., 2009), as it has been well studied and compared against other approaches (Turney and Mohammad, 2015). Formally, *BalAPinc* is the geometric mean of a symmetric similarity score, *LIN* (Lin, 1998), and an asymmetric score, *APinc*. Given a directional hypernym pair $(u \rightarrow v)$,

$$BalAPinc(u \rightarrow v) = \sqrt{LIN(u,v) \cdot APinc(u \rightarrow v)} \tag{4.1}$$

Assume we are given ranked feature lists FV_u and FV_v for words u and v respectively. Let $w_u(f)$ denote the weight of a particular feature f in FV_u. *LIN* is defined as

$$LIN(u,v) = \frac{\sum_{f \in FV_u \cap FV_v} [w_u(f) + w_v(f)]}{\sum_{f \in FV_u} w_u(f) + \sum_{f \in FV_v} w_v(f)} \tag{4.2}$$

APinc is a modified asymmetric version of the Average Precision metric used in Information Retrieval:

$$APinc(u \rightarrow v) = \frac{\sum_{r=1}^{|FV_u|} [P(r, FV_u, FV_v) \cdot rel'(f_r)]}{|FV_u|} \tag{4.3}$$

where,

$$P(r, FV_u, FV_v) = \frac{|\text{\# features of } v \text{ in top } r \text{ features of } u\ |}{r}$$

$$rel'(f) = \begin{cases} 1 - \frac{rank(f, FV_u)}{|FV_u|+1} & \text{if} f \in FV_u \\ 0 & \text{otherwise} \end{cases}$$

Thus, to use *BalAPinc* for cross-lingual hypernymy identification, we need a ranked list of features that capture information about the context of words in two languages. In the monolingual case, features are dimensions in a distributional semantic space. For the cross-lingual task, we need to represent words in two languages in the same space, or in spaces with a one-to-one mapping between dimensions.

4.3 BiSparse-Dep: Sparse Bilingual Word Representations using Dependency Contexts

There is a wealth of existing methods for learning representations that capture context of words in two different languages in the literature (Hermann and Blunsom, 2013; Luong et al., 2015; Upadhyay et al., 2016, *inter alia*). However, they have been evaluated on tasks that do not require much semantic analysis, such as bilingual lexicon induction or document categorization. In contrast, detecting hypernymy requires the ability to capture more subtle semantic distinctions. This requires bilingual representations to capture both the full range of word contexts observed in original language texts, as well as cross-lingual correspondences from translations.

We propose a new model that uses *sparse non-negative embeddings* to represent word contexts as interpretable dimensions, and facilitate context comparisons across languages. This is an instance of sparse coding, which consists of modeling data vectors as sparse linear combinations of basis elements. In contrast to dimensionality reduction techniques such as PCA, the learned basis vectors need not be orthogonal, which gives more flexibility to represent the data (Mairal et al., 2009). These models have been introduced as word representations in monolingual settings with the goal of obtaining interpretable, cognitively-plausible representations (Murphy et al., 2012) .

Additionally, work in monolingual setting has established that using dependency contexts to represent words (instead of window-based contexts) improves hy-

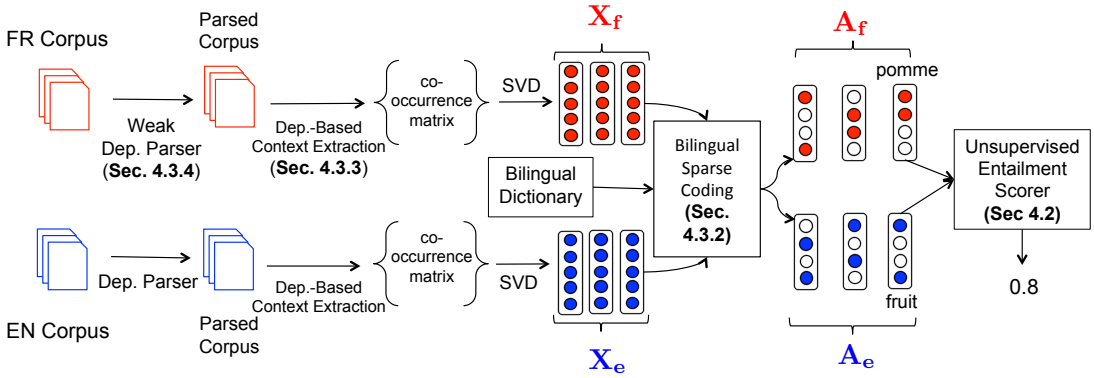

Figure 4.1: The BiSparse-Dep approach, which learns sparse bilingual embeddings using dependency based contexts. The resulting sparse embeddings, together with an unsupervised entailment scorer, can detect hypernyms across languages (e.g., *pomme* is a *fruit*).

pernymy detection, as dependency contexts capture richer information about both position and syntax (Levy and Goldberg, 2014; Lin, 1998; Shwartz et al., 2017). Motivated by this observation, we use syntactic contexts in order to represent words.

Figure 4.1 shows an overview of the end-to-end pipeline of our approach, which we call BiSparse-Dep. We first describe our generic framework for generating sparse bilingual embeddings (Section 4.3.2), and then describe how we extract dependency based contexts (Section 4.3.3). We also discuss how to extract such contexts in the absence of a treebank in the language (Section 4.3.4) using a (weak) dependency parser trained on related languages. The resulting sparse bilingual embeddings that are used with a unsupervised entailment scorer to predict hypernymy for cross-lingual word pairs, as described earlier (Section 4.2).

4.3.1 Review: Learning Monolingual Sparse Representations

Before introducing our novel bilingual formulation, we review the monolingual models for learning sparse representations. Previous work on obtaining sparse

monolingual representations is based on a variant of the Nonnegative Matrix Factorization problem (Faruqui et al., 2015; Murphy et al., 2012). Given a matrix $\mathbf{X}$ containing v dense word representations arranged row-wise, sparse representations for the v words can be obtained by solving the following optimization problem.

$$\begin{aligned} \arg\min_{\mathbf{A},\mathbf{D}} \quad & \sum_{i=1}^{v} ||\mathbf{A}_i\mathbf{D}^T - \mathbf{X}_i||_2^2 + \lambda||\mathbf{A}_i||_1 \\ \text{such that} \quad & \mathbf{A} \geq \mathbf{0} \qquad \|\mathbf{D}_i\|_2^2 \leq 1 \end{aligned} \tag{4.4}$$

The first term in the objective (Equation 4.4) factorizes the dense representation matrix $\mathbf{X}$ into two matrices, $\mathbf{A}$ and $\mathbf{D}$ such that the l_2 reconstruction error is minimized. The second term is an l_1 regularizer on $\mathbf{A}$ which encourages sparsity, where the level of sparsity is controlled by the λ hyperparameter. This, together with the non-negativity constraint, helps in obtaining sparse and interpretable representations in $\mathbf{A}$ since non-negativity has been shown to correlate with interpretability. The objective function on its own is degenerate since it can be trivially optimized by making the entries of $\mathbf{D}$ arbitrarily large and choosing corresponding small values as entries of $\mathbf{A}$. To avoid this, an additional l_2 constraint is imposed on $\mathbf{D}$.

4.3.2 BiSparse: Learning Sparse Bilingual Embeddings

Having seen how sparse representations can be obtained in monolingual settings, we now describe our novel formulation for obtaining such representations in bilingual settings using two sources of information.

- **Monolingual distributional representations independently learned from large amounts of text in each language**: We denote them as two input matrices, $\mathbf{X_e}$ and $\mathbf{X_f}$, of respective sizes $v_e \times n_e$ and $v_f \times n_f$. Each row in $\mathbf{X_e}$ is the representation of a particular word in the first language, $\boldsymbol{e}$, obtained using the contexts it appears in. Similarly, $\mathbf{X_f}$ contains word representations for the other language $\boldsymbol{f}$.

- **Cross-lingual correspondences that enable comparison across languages**: We define a "score" matrix $\mathbf{S}$ of size $v_e \times v_f$, which captures high-confidence correspondences between the vocabularies of the two languages. There are many ways of defining $\mathbf{S}$. As a starting point, we define each row of $\mathbf{S}$ as a one-hot vector that identifies the word in $\boldsymbol{f}$ that is most frequently aligned with the $\boldsymbol{e}$ word for that row in a large parallel corpus. This allows a many-to-one mapping from $\boldsymbol{e}$ to $\boldsymbol{f}$, which captures translation ambiguity by allowing multiple words in $\boldsymbol{e}$ to be aligned to the same word in $\boldsymbol{f}$.

Given this information, our BiSparse model solves the following optimization problem to obtain sparse bilingual representations.

$$\begin{aligned}\underset{\mathbf{A_e},\mathbf{D_e},\mathbf{A_f},\mathbf{D_f}}{\arg\min} & \sum_{i=1}^{v_e} \frac{1}{2}||\mathbf{A_e}_i{\mathbf{D_e}}^{\mathrm{T}} - \mathbf{X_e}_i||_2^2 \ + \lambda_e||\mathbf{A_e}_i||_1 \\ & + \sum_{j=1}^{v_f} \frac{1}{2}||\mathbf{A_f}_j{\mathbf{D_f}}^{\mathrm{T}} - \mathbf{X_f}_j||_2^2 \ + \lambda_f||\mathbf{A_f}_j||_1 \\ & + \sum_{i,j} \frac{1}{2}\lambda_x \mathbf{S}_{ij}||\mathbf{A_e}_i - \mathbf{A_f}_j||_2^2 \\ \text{such that} \quad & \mathbf{A_k} \geq \mathbf{0} \qquad \|\mathbf{D_k}_i\|_2^2 \leq 1 \qquad \mathbf{k} \in \{\mathbf{e}, \mathbf{f}\}\end{aligned} \tag{4.5}$$

The first two rows and the constraints in Equation 4.5 can be understood as in Equation 4.4—they encourage sparsity in word representations for each language. The third row imposes bilingual correspondence constraints, weighted by the regularizer λ_x. These constraints encourage words in $\boldsymbol{e}$ and $\boldsymbol{f}$ that are strongly aligned according to $\mathbf{S}$ to have similar representations.

4.3.3 BiSparse-Dep: Inducing Dependency Based Contexts

The BiSparse framework requires contextual representations of words in the two languages as inputs $\mathbf{X_e}$ and $\mathbf{X_f}$. One way to construct these is by representing a target word using its window-context, *i.e.* treating the words that appear to the left and the right of the target word as the context. However, work in monolingual settings has established that syntactic context is more valuable than window-context when it comes to hypernymy detection, as it captures both positional and syntactic information, as opposed to window-based context, which only contains positional information. Moreover, syntactic contexts capture functional similarity (*e.g.* lion-cat) rather than the topical similarity (*e.g.* lion-zoo) that window-based contexts capture, and the former is more essential to hypernymy identification. Thus, our final model BiSparse-Dep, uses syntactic contexts extracted from dependency graphs to represent words.

Given a dependency graph, the context of a word can be described in multiple ways using its syntactic neighborhood in the graph. For instance, in Figure 4.2, we describe the context for a target word (*traveler*) in the following two ways:

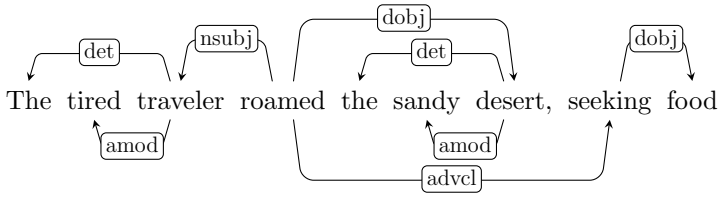

Figure 4.2: Example Dependency Tree.

- FULL context (Baroni and Lenci, 2010; Levy and Goldberg, 2014; Padó and Lapata, 2007): Children and parent words, concatenated with the label and direction of the relation (*e.g.* *roamed#nsubj*$^{-1}$ and *tired#amod* are contexts for *traveler*).

- JOINT context (Chersoni et al., 2016b): Parent concatenated with each of its siblings (*e.g.* *roamed#desert* and *roamed#seeking* are contexts for *traveler*).

These two contexts exploit different amounts of syntactic information—FULL requires labeled parses, unlike JOINT. JOINT combines parent and sibling information, while FULL treats them distinctly. Both encode direction of the dependency into the context through label direction and sibling-parent relations respectively.

4.3.4 Dependency Contexts without a Treebank

Using dependency contexts in multilingual settings may not always be possible, as dependency treebanks are not available for many languages. However, related languages show common syntactic structure that can be transferred to the original language of interest, with *delexicalized parsing* being one common approach (McDonald et al., 2011; Zeman and Resnik, 2008, *inter alia).*

To extract contexts for BISPARSE-DEP for a language without a treebank,

we train a delexicalized parser using treebanks of related languages, where the word form based features are turned off, so that the parser is trained on purely non-lexical features (*e.g.* POS tags). More sophisticated techniques for transferring syntactic knowledge have been proposed (Ammar et al., 2016; Rasooli and Collins, 2017), but we prioritize simplicity and show that a simple delexicalized parser is effective.

4.3.5 Optimization of BiSparse formulation

While we have described our formulation for obtaining sparse representations (Equation 4.5), and the inputs to the model, we have not described how we can solve the formulation to obtain the expected output of sparse embeddings. Equation 4.5 defines a non-differentiable, non-convex optimization problem and finding the globally optimally solution is not feasible. However, various methods used to solve convex problems work well in practice. We use *Forward Backward Splitting*, a proximal gradient method for which an efficient generic solver, FASTA, is available (Goldstein et al., 2014).[2] FASTA (Fast Adaptive Shrinkage / Thresholding Algorithm) is designed to minimize functions of the form $f(Ax) + g(x)$, where f is a differentiable function, g is a function (possibly non-differentiable) for which we can calculate the proximal operator, and A is a linear operator. For the objective function in our model, the l_1 terms form g and the l_2 terms form f.

[2] `https://github.com/tomgoldstein/fasta-matlab`

4.4 Crowd-Sourcing Annotations

To measure the accuracy of approaches for identifying cross-lingual hypernymy, we collect and release reliable and high-quality testbeds spanning four languages from distinct families—French (Fr), Russian (Ru), Arabic (Ar) and Chinese (Zh)—paired with English. Lack of available datasets to evaluate models of hypernymy detection across multiple languages aggravates the need for a high quality test bed. While ontologies like Open Multilingual WordNet (OMW) (Bond and Foster, 2013) and BabelNet (Navigli and Ponzetto, 2012) contain cross-lingual links, these resources are semi-automatically generated and hence contain noisy edges. Below, we describe how we pool candidates from such resources, and crowdsource high-quality annotations to create the testbeds.

4.4.1 Annotation Setup

We collect evaluation datasets using Crowdflower.[3] The annotation task for a particular target language requires annotators to be fluent in both the target language and English. To ensure only fluent speakers perform the task, for each target language, we provide task instructions in that language itself. Also, we restrict the task to annotators verified by Crowdflower to have those language skills. Finally, annotators also need to pass a quiz based on a small amount of gold standard data to gain access to the task.

To begin the annotation process, we first obtain candidate pairs using hyper-

[3] `http://crowdflower.com`

nymy edges across languages from OMW and BabelNet, along with translations from monolingual hypernymy datasets (Baroni and Lenci, 2011; Baroni et al., 2012; Kotlerman et al., 2010). Annotators choose between three options for each candidate pair (p_f, q_e), where p_f is a target language word and q_e is a English word : "p_f is a kind of q_e", "q_e is a part of p_f" and "none of the above". Word pairs labeled with the first option are considered as positive examples while those labeled as "none of the above" are considered as negative. The second option is included to filter out meronymy examples that are part of the noisy pool. We leave it to the annotator to infer whether the relation holds between any senses of p_f or q_e, if either of them are polysemous. We collect more negative pairs than positive, but down-sample the negative examples to keep a balanced dataset for ease of evaluation.

For every candidate hypernym pair (p_f, q_e), we also ask annotators to judge its reversed and translated *hyponym* pair (q_f, p_e). For instance, if (*citron*, *food*) is a hypernym candidate, we also show annotators (*aliments*, *lemon*) which is a potential hyponym candidate (potential, because translation need not preserve semantic relations). The purpose of presenting the hyponym pair, (q_f, p_e), is two-fold. First, it emphasizes the directional nature of the task. We want annotators to accurately identify that *food* is a hypernym of *citron*, but not the other way around. Second, it identifies hyponym pairs, which we use as negative examples. The hyponym pairs are challenging since differentiating them from hypernyms requires detecting asymmetry.

Each pair is judged by at least five annotators, and judgments with 80% agreement are considered for the final dataset. This is a stricter condition than certain

Language pair	#crowdsourced	#pos (= #neg)
French-English	2115	763
Russian-English	2264	706
Arabic-English	2144	691
Chinese-English	2165	806

Table 4.1: Statistics of our crowd-sourced test bed. #pos (#neg) denote positives (negatives) in the evaluation set. We down-sample negatives to have a balanced evaluation set.

monolingual hypernymy datasets—for instance, EVALution (Santus et al., 2015)—where agreement by 3 annotators is deemed sufficient. Inter-annotator agreement measured using Fleiss' Kappa (Fleiss, 1971) is 58.1 (French), 53.7 (Russian), 53.2 (Arabic) and 55.8 (Chinese). This indicates moderate agreement, on par with agreement obtained on related fine-grained semantic tasks (Pavlick et al., 2015). We cannot compare with monolingual hypernymy annotator agreement as, to the best of our knowledge, such numbers are not available for existing test sets. Dataset statistics are shown in Table 4.1.

We observe that annotators were able to agree on pairs containing polysemous words where hypernymy holds for some sense. For instance, for the French-English pair (*avocat*, *professional*), the French word *avocat* can either mean *lawyer* or *avocado*, but the pair is annotated as a positive example. Hence, we leave it to the annotators to handle polysemy by choosing the most appropriate sense.

4.4.2 Two Evaluation Test Sets

To verify if the crowdsourced hyponyms are challenging negative examples we create two evaluation sets. Both share the (crowdsourced) positive examples, but

differ in their negatives:

- HYPER-HYPO: negative examples are the crowdsourced hyponyms.
- HYPER-COHYPO: negative examples are *cohyponyms* drawn from OMW.

Cohyponyms are words sharing a common hypernym. For instance, *bière* ("beer" in French) and *vodka* are cohyponyms since they share a common hypernym in *alcool*/*alcohol*. We choose cohyponyms for the second test set because: (a) They require differentiating between similarity (a symmetric relation) and hypernymy (an asymmetric relation). For instance, *bière* and *vodka* are highly similar but they do not have a hypernymy relationship. (b) Cohyponyms are a popular choice of negative examples in many entailment datasets (Baroni and Lenci, 2011).

4.5 Experimental Setup

4.5.1 Data and Evaluation Setup

Training BISPARSE-DEP requires a dependency parsed monolingual corpus, and a translation matrix for jointly aligning the monolingual vectors. We compute the translation matrix using word alignments derived from parallel corpora (see corpus statistics in Table 4.2). While we use parallel corpora to generate the translation matrix to be comparable to baselines (Section 4.5.2), we can obtain the matrix from any bilingual dictionary.

The monolingual corpora are parsed using Yara Parser (Rasooli and Tetreault, 2015), trained on the corresponding treebank from the Universal Dependency Tree-

bank (McDonald et al., 2013) (UDT-v1.4). Yara Parser was chosen as it is fast, and competitive with state-of-the-art parsers (Choi et al., 2015). The monolingual corpora was POS-tagged using TurboTagger (Martins et al., 2013). We induce dependency contexts for words by first thresholding the language vocabulary to the top 50,000 nouns, verbs and adjectives. A co-occurrence matrix is computed over this vocabulary using the context types in Section 4.3.3.

Inducing Dependency Contexts The entries of the word-context co-occurrence matrix are re-weighted using Positive Pointwise Mutual Information (Bullinaria and Levy, 2007). The resulting matrix is reduced to 1000 dimensions using SVD (Golub and Kahan, 1965).[4] These vectors are used as $\mathbf{X_e}$ and $\mathbf{X_f}$ in Equation 4.5 to generate 100 dimensional sparse bilingual vectors.

Evaluation We use accuracy as our evaluation metric, as it is easy to interpret when the classes are balanced (Turney and Mohammad, 2015). Both evaluation datasets—Hyper-Hypo and Hyper-Cohypo—are split into 1:2 dev/test splits. *BalAPinc* has two tunable parameters - 1) a threshold that indicates the *BalAPinc* score above which all examples are labeled as positive, 2) the maximum number of features to consider for each word. We use the tuning set to tune these two parameters as well as the various hyperparameters associated with the models.

[4]Chosen based on preliminary experiments with {500,1000,2000,3000} dimensional vectors for En-Fr.

Language	Parallel Data	#sent.	Monolingual Data	#sent.
English	–	–	Wackypedia (Baroni et al., 2009)	43M
Arabic	ISI (LDC2007T08), NewsCommentary, Wikipedia (Tiedemann, 2012)	1.1M	Arabic Gigaword 3.0 (LDC2007T40)	17M
Chinese	FBIS (LDC2003E14)	9.5M	Chinese Gigaword 5.0 (LDC2011T13)	58M
French	Europarl (Koehn, 2005), NewsCommentary♦, Wikipedia (Tiedemann, 2012)	2.7M	Wikipedia♣	20M
Russian	Yandex-1M♠	1.6M	Wikipedia♣	22M

♦ = `www.statmt.org/wmt15/training-parallel-nc-v10.tgz`
♣ = `dumps.wikimedia.org/xxwiki/20161201/`
♠ = `translate.yandex.ru/corpus`

Table 4.2: Training data statistics for different languages. Note that while we use parallel corpora for computing translation dictionaries, our approach does not require it, and can work with any bilingual dictionary.

4.5.2 Contrastive Approaches

We compare our BiSparse-Dep embeddings with the following approaches:

Mono-Dep **(Translation baseline)** For each word pair (p_f, q_e) in the test data, we translate p_f to English using the most common translation in the translation matrix. Hypernymy is then determined using sparse, dependency based embeddings in English.

BiSparse-Lex **(Window context)** Instead of using dependency context, this approach generates sparse, cross-lingual embeddings using a window based context

allowing us to directly compare the relative importance of dependency contexts and window based contexts.

BIVEC+ **(Window context)** Our extension of the BIVEC model of Luong et al. (2015). BIVEC generates dense, cross-lingual embeddings using window based context, by substituting aligned word pairs within a window in parallel sentences. By default, BIVEC only trains using parallel data, so we initialize it with monolingually trained window based embeddings to ensure fair comparison.

CL-DEP **(Dependency context)** The model from Vulić (2017), which induces dense, dependency based cross-lingual embeddings by translating syntactic word-context pairs using the most common translation, and jointly training a `word2vecf` model for both languages.[5] Vulić (2017) showed improvements for word similarity and bilingual lexicon induction. We report the first results using CL-DEP on this task. By comparing with dense representations induced by BIVEC and CL-DEP, we can identify the importance of our sparse coding framework.

4.5.3 Evaluating Robustness of BISPARSE-DEP

We investigate how robust BISPARSE-DEP is when exposed to data scarce settings. Evaluating on a truly low resource language is complicated by the difficulty of obtaining an evaluation dataset for such a language. Therefore, we simulate such settings for the languages in our dataset in multiple ways.

[5] `bitbucket.org/yoavgo/word2vecf/`

No Treebank If a treebank is not available for a language, dependency contexts have to be induced using treebanks from other languages (Section 4.3.4), which can affect the quality of the dependency-based embeddings. To simulate this, we train a delexicalized parser for all four languages. We use treebanks from Slovenian, Ukrainian, Serbian, Polish, Bulgarian, Slovak and Czech (40k sentences) for training the Russian parser, and treebanks from English, Spanish, German, Portuguese, Swedish and Italian (66k sentences) for training the French parser. At the time of this work, UDT did not have languages in the same family as Arabic or Chinese, so for the sake of completeness, we train Arabic and Chinese parsers on delexicalized treebanks of the language itself. After delexicalized training, the Labeled Attachment Score (LAS) on the UDT test set dropped for all languages (76.6% to 60.0% for Russian, 83.7% to 71.1% for French, 76.3% to 62.4% for Arabic, and 80.3% to 53.3% for Chinese). The monolingual corpora are then parsed with these weaker parsers, and co-occurrences and dependency contexts are computed as before.

Subsampling Monolingual Data To simulate low-resource behavior along another axis, we subsample the monolingual corpora used by BiSparse-Dep to induce the monolingual vectors, $\mathbf{X_e}$ and $\mathbf{X_f}$. Specifically, we learn $\mathbf{X_e}$ and $\mathbf{X_f}$ using progressively smaller corpora.

Quality of Bilingual Dictionary We study the impact of the quality of the bilingual dictionary used to create the translation matrix $\mathbf{S}$. This experiment involves using increasingly smaller parallel corpora to induce the translation dictionary.

Model	Ru-En	Zh-En	Ar-En	Fr-En	Average
Translation Baseline					
Mono-Dep	50.1	52.3	51.8	54.5	52.2
Window Based Contexts					
BiSparse-Lex	56.6	53.7	50.9	52.0	53.3
Bivec+	55.8	52.0	51.5	53.4	53.2
Dependency Based Contexts					
Cl-Dep	**60.2**	54.4	**56.7***	53.8	**56.3**
BiSparse-Dep (Full)	59.0	55.9	52.6	56.6	56.0
BiSparse-Dep (Joint)	53.8	**57.0***	52.4	**59.9***	55.8
BiSparse-Dep (Unlabeled)	55.9	51.2	53.3	55.9	54.1

Table 4.3: Comparing the different approaches from Section 4.5.2 with our BiSparse-Dep approach on Hyper-Hypo (random baseline = 0.5). **Bold** denotes the best score for each language, and the * on the best score indicates a statistically significant ($p < 0.05$) improvement over the next best score, using McNemar's test (McNemar, 1947). Dependency based models largely outperform window based models, and all BiSparse-Dep models outperform translation baselines.

4.6 Experiments

Our experiments aim to answer the following questions — **(a)** Are dependency based embeddings superior to window based embeddings for identifying cross-lingual hypernymy? (Section 4.6.1) **(b)** Does directionality in the dependency context help cross-lingual hypernymy identification? (Section 4.6.2) **(c)** Are our models robust in data scarce settings (Section 4.6.3)? **(d)** Is the answer to **(a)** predicated on the choice of entailment scorer? (Section 4.6.4)?

Model	Ru-En	Zh-En	Ar-En	Fr-En	Average
Translation Baseline					
MONO-DEP	58.7	50.0	65.1	56.9	57.7
Window Based Contexts					
BISPARSE-LEX	**63.8**	55.8	65.8	63.2	62.2
BIVEC+	55.9	64.9	62.2	54.1	58.3
Dependency Based Contexts					
CL-DEP	56.2	62.7	63.1	61.0	60.0
BISPARSE-DEP (Full)	63.6	**67.3**	**66.8***	**66.7***	**66.1**
BISPARSE-DEP (Joint)	60.6	63.6	65.9	64.9	63.8
BISPARSE-DEP (Unlabeled)	58.6	66.7	62.4	61.5	62.4

Table 4.4: Comparing the different approaches from Section 4.5.2 with our BISPARSE-DEP approach on HYPER-COHYPO (random baseline = 0.5). **Bold** denotes the best score for each language, and the * on the best score indicates a statistically significant ($p < 0.05$) improvement over the next best score, using McNemar's test (McNemar, 1947). BISPARSE-DEP models continue to outperform window based models and the translation baseline on an average.

4.6.1 Dependency v/s Window Contexts

We compare the performance of models described in Section 4.5.2 with the BISPARSE-DEP (FULL and JOINT) models. We evaluate the models on the two test splits described in Section 4.4.2, *i.e.* HYPER-HYPO and HYPER-COHYPO.

Hyper-Hypo Results First, results in Table 4.3 highlight the benefit of cross-lingual modeling (as opposed to translation). Almost all models (except CL-DEP on French) outperform the translation baseline. Among dependency based models, BISPARSE-DEP (FULL) and CL-DEP consistently outperform both window models, while BISPARSE-DEP (JOINT) outperforms them on all except Russian. BISPARSE-DEP (JOINT) is best overall for two languages (French and Chinese), CL-DEP for

one (Arabic), with no statistically significant differences between BiSparse-Dep (Joint) and Cl-Dep for Russian. This confirms that dependency context is more useful than window context for cross-lingual hypernymy detection.

Hyper-Cohypo Results The trends observed on Hyper-Hypo also hold on Hyper-Cohypo, *i.e.* dependency based models continue to outperform window based models (Table 4.4). Overall, BiSparse-Dep (Full) performs best in this setting, followed closely by BiSparse-Dep (Joint). This suggests that the sibling information encoded in Joint is useful to distinguish hypernyms from hyponyms (Hyper-Hypo results), while the dependency labels encoded in Full help to distinguish hypernyms from co-hyponyms. All models also improve significantly on the Hyper-Cohypo set, suggesting that discriminating hypernyms from cohyponyms is easier than discriminating them from hyponyms.

While the BiSparse-Dep models generally perform better than window models on both test sets, Cl-Dep is not as consistent (*e.g.* it is worse than the best window model on Hyper-Cohypo). As shown by Turney and Mohammad (2015), *BalAPinc* is designed for sparse embeddings and is likely to perform poorly with dense embeddings. This explains the relatively inconsistent performance of Cl-Dep.

Finally, besides establishing the challenging nature of our crowd-sourced set, the experiments on Hyper-Cohypo and Hyper-Hypo also demonstrate the ability of the BiSparse-Dep models to discriminate between different lexical-semantic relations (viz. hypernymy and cohyponymy) in a cross-lingual setting.

4.6.2 Ablating Directionality in Context

The context described by the FULL and JOINT BISPARSE models encodes directional information (Section 4.3.3) either in the form of label direction (FULL), or using sibling information (JOINT). Does such directionality in the context help to capture the asymmetric relationship inherent to hypernymy? To answer this, we evaluate a third BISPARSE-DEP model which uses UNLABELED dependency contexts. This is similar to the FULL context, except we do not concatenate the label of the relation to the context word (parent or children). For instance, for *traveler* in Figure 4.2, contexts will be *roamed* and *tired*.

Experiments on both HYPER-HYPO and HYPER-COHYPO (bottom row, Tables 4.3 and 4.4) highlight that directional information is indeed essential. UNLABELED almost always performs worse than FULL and JOINT, and in many cases worse than even window based models.

4.6.3 Evaluating Robustness of BISPARSE-DEP

No Treebank We run experiments (Table 4.5) for all languages with a version of BISPARSE-DEP that use the FULL context type for both English and the non-English (target) language, but the target language contexts are derived from a corpus parsed using a delexicalized parser (Section 4.5.3). This model compares favorably on all language pairs against the best window based and the best dependency based model. In fact, it almost consistently outperforms the best window based model by several points, and is only slightly worse than the best dependency-based model.

Model	Ru-En	Zh-En	Ar-En	Fr-En	Average
Hyper-Hypo					
Best Window	56.6	53.7	51.5	53.4	53.8
Delexicalized	59.1*	55.1*	54.6*	56.1*	56.2
Best Dependency	60.2	57.0*	56.7*	59.9*	58.5
Hyper-Cohypo					
Best Win.	63.8	64.9	65.8	63.2	64.4
Delexicalized	59.4	65.7*	67.5*	66.3*	64.7
Best Dependency	63.6*	67.3*	66.8*	66.7	66.1

Table 4.5: The delexicalized model is competitive with the best dependency based and the best window based models on both test sets. For each dataset, * indicates a statistically significant ($p < 0.05$) improvement over the next best model in that column, using McNemar's test (McNemar, 1947).

Further analysis reveals that the strong performance of the delexicalized model is due to the relative robustness of the delexicalized parser on frequent contexts in the co-occurrence matrix. Specifically, we find that in French and Russian, the most frequent contexts were derived from `amod`, `nmod`, `nsubj` and `dobj` edges (together they make up at least 70% of the contexts). For instance, the `nmod` edge appears in 44% of Russian contexts and 33% of the French contexts. The delexicalized parser predicts both the label and direction of the `nmod` edge correctly with an F1 of 68.6 for Russian and 69.6 for French. In contrast, a fully-trained parser achieves a F1 of 76.7 for Russian and 76.8 for French for the same edge.

Small Monolingual Corpus In Figure 4.3, we use increasingly smaller monolingual corpora (10%, 20%, 40%, 60% and 80%) sampled at random to induce the monolingual vectors for BiSparse-Dep (Full) model. Trends indicate that BiSparse-Dep models that use only 40% of the original data remain competitive with the BisSparse-Lex model that has access to the full data. Robust perfor-

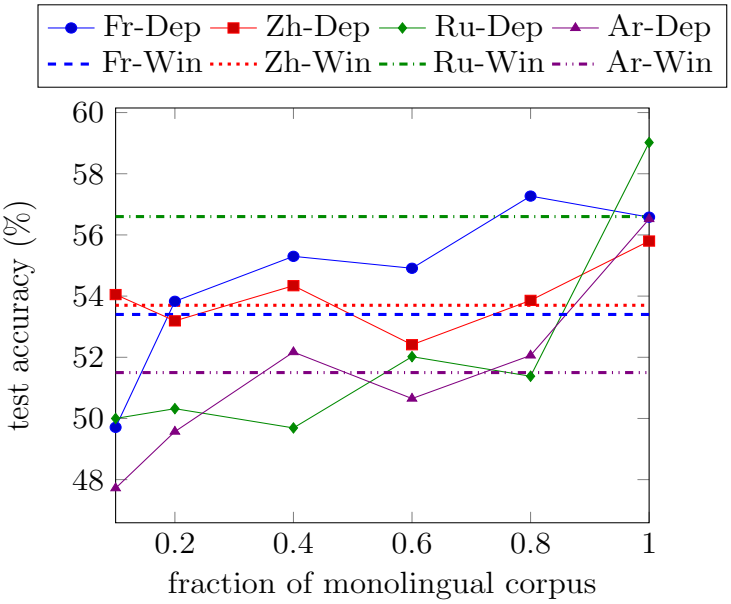

Figure 4.3: BiSparse-Dep outperforms the best window based model on Hyper-Hypo, even with about 40% of the monolingual corpora, for most languages.

mance with smaller monolingual corpora is helpful since large monolingual corpora are not always easily available.

Quality of Bilingual Dictionary Bilingual dictionaries derived from smaller amounts of parallel data are likely to be of lower quality than those derived from larger corpora. To analyze the impact of dictionary quality on BiSparse-Dep (Full), we use increasingly smaller parallel corpora to induce bilingual dictionaries used as the score matrix **S** (Section 4.3.2). We use the top 10%, 20%, 40%, 60% and 80% sentences from the parallel corpora. Trends in Figure 4.4 show that even with a lower quality dictionary, BiSparse-Dep is more accurate than BiSparse-Lex.

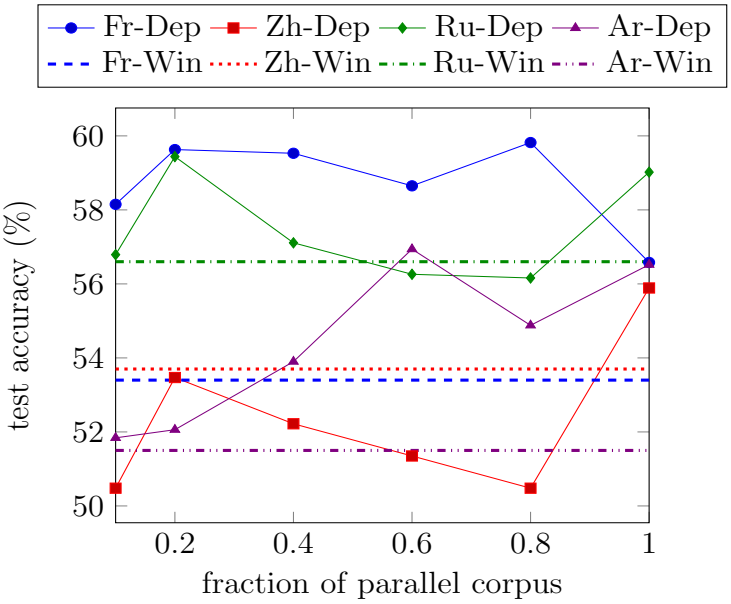

Figure 4.4: BiSparse-Dep outperforms the best window based model on Hyper-Hypo with increasingly lower quality dictionaries, for most languages.

4.6.4 Choice of Entailment Scorer

While BiSparse-Dep uses *BalAPinc* to score hypernyms, how robust is it to other functions motivated by different linguistic hypotheses? To answer this, we change the hypernymy scorer from *BalAPinc* to *SLQS* (Santus et al., 2014) and redo experiments (Section 4.6.1). *SLQS* is based on the **distributional informativeness hypothesis**, which states that hypernyms are less "informative" than hyponyms, as they occur in more general contexts. The informativeness E_u of a word u is defined as the median entropy of its top N dimensions, $E_u = median\ {}^{N}_{k=1} H(c_k)$, where $H(c_i)$ is the entropy of dimension c_i. The *SLQS* score for a pair (u, v) is the relative difference in entropies,

$$SLQS(u \rightarrow v) = 1 - \frac{E_u}{E_v} \tag{4.6}$$

Recent work (Shwartz et al., 2017) has found *SLQS* to be more successful than other metrics in monolingual hypernymy detection.

The trends observed in these experiments are consistent with those in Section 4.6.1—both BiSparse-Dep models still outperform window-based models. Also, the delexicalized version of BiSparse-Dep outperforms the window-based models, showing that the robust behavior demonstrated in Section 4.6.3 is also invariant across metrics. We also find that using *BalAPinc* leads to better results than *SLQS*. For both BiSparse-Dep models, *BalAPinc* wins across the board for two languages (Russian and Chinese), and wins half the time for the other two languages.

4.7 Summary

The focus of this chapter was on the task of identifying hypernymy relations between words across languages. We introduce this task in order to provide a principled way of characterizing differences in word meaning across languages. Motivated by the distributional inclusion hypothesis, we introduced BiSparse-Dep, a new unsupervised approach for identifying cross-lingual hypernymy. BiSparse-Dep uses sparse bilingual word representations learned from both dependency-based co-occurrence patterns in monolingual corpora and bilingual correspondences from parallel text. We showed that BiSparse-Dep is superior for the cross-lingual hypernymy detection task, when compared to standard window based models and a translation baseline. Further analysis showed that BiSparse-Dep is robust to various low-resource settings. In principle, BiSparse-Dep can be used for any lan-

guage that has a bilingual dictionary with English and a related language with a treebank. We also introduced crowd-sourced cross-lingual hypernymy datasets for four languages and make them publicly avaiable for future evaluations.

The accuracy of BiSparse-Dep reveals that hypernymy identification can still be further improved, especially in the harder setting where it has to be distinguished from the inverse hyponymy relation. In monolingual settings, improvements have been obtained using supervised methods, but direct supervision is unrealistic to assume in cross-lingual scenarios. Instead, methods that learn from supervised data in a high-resource language (Glavaš and Vulić, 2019; Vulić et al., 2019) provide a reasonable alternative.

Chapter 5: Weakly Supervised Identification of Cross-lingual Semantic Relations

While hypernymy is one way of characterizing differences in meaning across languages, words across languages can be related in many different ways. In this chapter, we expand our scope to the broader challenge of simultaneously classifying between multiple relations. Given a word pair (water, पय), the classification task is to select one of the five entailment classes (Figure 5.1) defined under the natural logic framework of MacCartney and Manning (2009). The challenges associated with hypernymy detection in the previous chapter still remain — we cannot assume that labeled examples exist for all language pairs and learning from English labeled examples is complicated by translation ambiguity, and by semantic relations not being preserved by translation, as illustrated by Figure 5.1.

We introduce BiLexNet, a neural classifier for semantic relations based on cross-lingual distributional and path-based features inspired by the monolingual LexNet model (Shwartz and Dagan, 2016b,c) (Section 5.1). LexNet achieved the highest performance (45 F1) among participating teams on the CogALex-V shared task on identification of semantic relations (Santus et al., 2016) without ontologies or structured information. We adapt LexNet to make cross-lingual predictions by

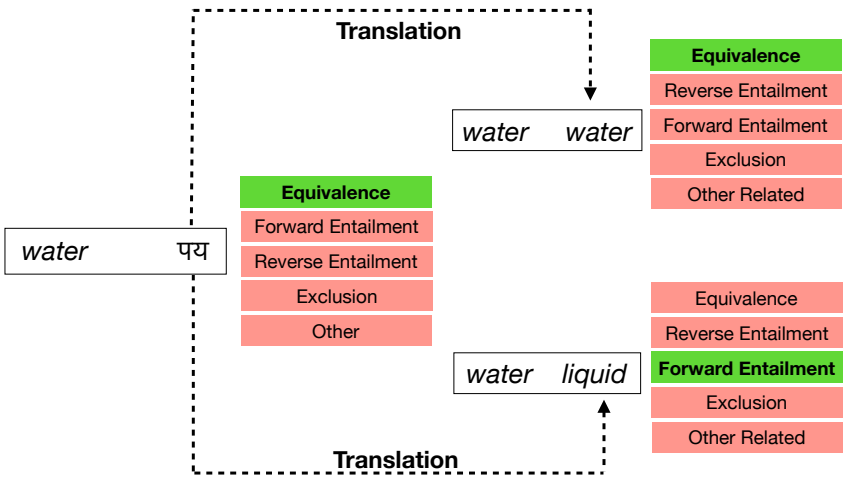

Figure 5.1: On the left, we illustrate cross-lingual semantic relation classification: given the pair (*water*, **पय**) as input, the task is to select the `Equivalence` class (in bold/green) from the five possible relations. On the right, we show that semantic relations change by translation. **पय** translates to *liquid* and *water*, and their respective semantic relations with *water* differ.

proposing to model cross-lingual relations using lexico-syntactic paths from both languages.

We then design a novel training procedure for BiLexNet that leverages weak supervision in the form of examples translated from English via a knowledge distillation technique guided by translation dictionaries (Hinton et al., 2015) (Section 5.2). Knowledge distillation has been proposed to compress a model with many parameters (the *teacher* model) to a model with fewer parameters (the *student* model). It has also been used successfully to learn mappings between languages (Nakashole and Flauger, 2017) or to transfer knowledge from models trained on one language to a different target language for text classification (Xu and Yang, 2017) and belief tracking (Chen et al., 2018a), in settings where the classification labels are translation invariant. This work adapts distillation to a setting where labels might change when samples are translated.

Finally, we collect and release MULTILEXREL, a crowdsourced benchmark to evaluate models for this task on a high-resource (English-Chinese) and a low-resource (English-Hindi) language pair (Section 5.3). Experiments show that BILEXNET substantially outperforms translation baselines and approaches the performance of a fully supervised English semantic relation classifier (Section 5.5).

5.1 BILEXNET: a Classifier for Cross-Lingual Semantic Relations

The task of classifying semantic relations is a multi-class classification problem, where the classes are the set of five semantic relations from Pavlick et al. (2015): `Equivalence` (X is the same as Y), `Forward Entail` (X is more specific than/a type of Y), `Backward Entail` (X is more general than/encompasses Y), `Exclusion` (X is mutually exclusive with/opposite to Y), and `Other` (X is not related or related in other ways to Y). We choose these relations as they have been useful in describing lexical relations between English paraphrases (Pavlick et al., 2015), and in downstream natural language inference systems (MacCartney and Manning, 2007, 2009).

Our classifier, BILEXNET, adapts the LEXNET English classifier (Shwartz and Dagan, 2016b,c) to cross-lingual settings. BILEXNET represents the input word pair (x, y) by a feature vector $\mathbf{v}_{xy}$, consisting of complementary distributional and path-based features i.e. $\mathbf{v}_{xy}$= $[\mathbf{v}_x;\mathbf{v}_y;\mathbf{v}_{paths(x,y)}]$. The *distributional semantic* properties of x and y are captured by bilingual word embeddings $\mathbf{v}_x$ and $\mathbf{v}_y$. $\mathbf{v}_{paths(x,y)}$ encodes *lexico-syntactic paths* that represent the relation between words x and y in con-

text (Hearst, 1992; Shwartz et al., 2016; Snow et al., 2005). For classification, $\mathbf{v}_{xy}$ is input to a multi-class classifier, parameterized as a feed-forward neural network with a single hidden layer.

$$\begin{aligned} \mathbf{l}_{out} &= \mathbf{W}_2 * \text{ReLU}(\mathbf{W}_1 * \mathbf{v}_{xy}) \\ \hat{l}_i &= \frac{\exp(l_{out,i})}{\sum_{j=1}^{k} \exp(l_{out,j})} \\ l_{pred} &= \arg\max_i \hat{l}_i \end{aligned} \tag{5.1}$$

$\mathbf{W}_1$ and $\mathbf{W}_2$ are the weights of the network, and the biases have been omitted for simplicity.

5.1.1 Cross-lingual Paths

In LexNet, a lexico-syntactic path is the *sequence* of edges that lead from x to y in the dependency tree of a sentence. Each edge contains the word and part-of-speech tag of the source node, the dependency label of the edge, and the edge direction between two subsequent nodes (see Figure 5.2 for an example). The vector representation of each edge is formed by concatenating the vectors of these four components. $\mathbf{v}_{paths(x,y)}$ is obtained by encoding the sequence of edges using an LSTM (Hochreiter and Schmidhuber, 1997).

In English, these paths are extracted from sentences where x and y co-occur. However, when x and y are in different languages, a new path definition is required. For a cross-lingual pair (x_e, y_f), we extract cross-lingual paths $\mathbf{v}_{paths(x_e,y_f)}$ from a

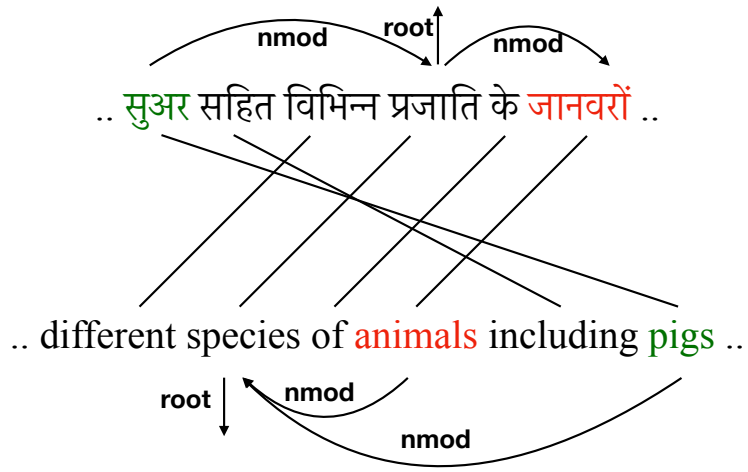

Figure 5.2: The English path between *animals* and *pigs* has three edges: [X/NOUN/nmod/>, species/NOUN/root/∧, and Y/NOUN/nmod/>]. The path between animals and सुअर is defined as a combination of the English path and the Hindi path between जानवरों and सुअर.

word-aligned parallel corpus (Figure 5.2). We first extract all parallel sentences which contain x_e on the source side and y_f on the target side. For each sentence, using word alignments, we can extract x_f, the target word aligned to x_e, and y_e the source word aligned to y_f. We then extract a path connecting the two word in the source sentence *i.e.* x_e and y_e. Similarly, we also extract a corresponding path connecting the two word in the target sentence *i.e.* x_f and y_f, since different languages can encode the same information differently due to structural divergences (Dorr, 1994). Thus, if the parallel corpus contains m sentence pairs where x_e occurs on the source side and y_f on the target side, we extract a total of $2m$ paths. All of the $2m$ paths paths are encoded using a single LSTM, and averaged to form $\mathbf{v}_{paths(x_e,y_f)}$.

Two special cases arise from this definition. First, a path can be a single alignment link if x_e and y_f are aligned to each other *i.e.* $x_f = y_f$ and $y_e = x_e$. Second, if no path is found in the corpus, $\mathbf{v}_{paths(x_e,y_f)}$ is set to the zero vector.

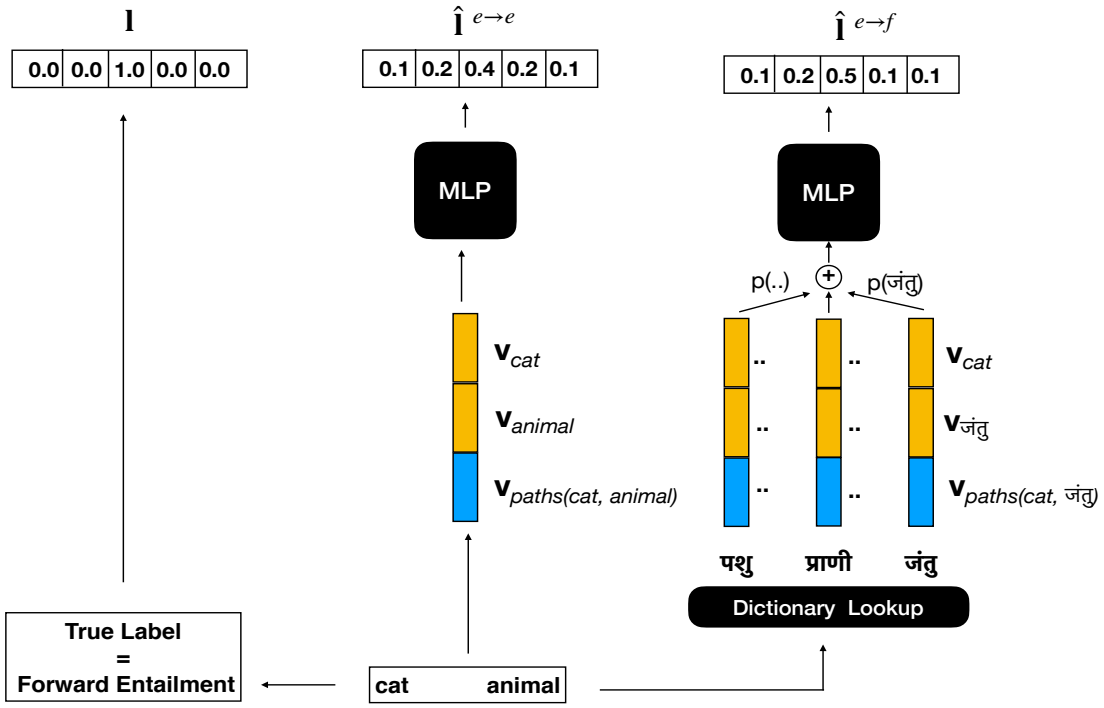

Figure 5.3: Illustration of weakly supervised training: For a given English example (*cat*, *animal*), we generate predictions $\hat{\mathbf{l}}^{e \to e}$ using the monolingual English teacher model. Simultaneously, we also generate predictions $\hat{\mathbf{l}}^{e \to f}$ using the cross-lingual *student* model after translating one of the two English words using a dictionary. The cross-lingual classifier attends to all translation candidates and predicts a class based on a weighted average of their features. The loss is defined as CROSS-ENTROPY($\hat{\mathbf{l}}^{e \to e}$, $\mathbf{l}$) + CROSS-ENTROPY($\hat{\mathbf{l}}^{e \to f}$, $\mathbf{l}$) + KL-DIVERGENCE($\hat{\mathbf{l}}^{e \to f}$, $\hat{\mathbf{l}}^{e \to e}$).

5.2 Weakly Supervised Training via Knowledge Distillation

Cross-lingual examples for fully supervised training of BILEXNET are hard to obtain: examples of relations such as synonymy or hypernymy can be derived from multilingual WordNets (Bond and Foster, 2013), but such resources are not available for many languages, and only cover a subset of semantic relations. Instead, we introduce a dictionary-guided variant of *knowledge distillation* to train BILEXNET. This procedure only relies on a set of monolingual labeled examples that are readily available for various lexical relations in English, and a translation dictionary that

maps words in the source language to the target language.

Our approach transfers knowledge from a monolingual *teacher model* to a cross-lingual *student model*. The **teacher model** is a monolingual LexNet model (say M_e) trained on the source-language examples $S = \{(x_{e;i}, y_{e;i}, \mathbf{l}_i)\}$. Here, $x_{e;i}$ and $y_{e;i}$ are a pair of words in the same language and $\mathbf{l}_i \in \mathbb{R}^c$ is a one-hot encoding of the relation between $x_{e;i}$ and $y_{e;i}$ (the number of possible relations is c). Given $(x_{e;i}, y_{e;i}, \mathbf{l}_i) \in S$, M_e is trained by minimizing the cross-entropy loss between the predicted output $\hat{\mathbf{l}}^{\,e \to e}$ and the gold label $\hat{\mathbf{l}}_i$:

$$L_1 = -\sum_{j=1}^{c} l_{ij} \log \hat{l}_j^{\,e \to e} \tag{5.2}$$

BiLexNet plays the role of the **student model** (denoted M_{ef}) and is trained to make predictions that agree with those of the teacher model. The student model is trained using **weak supervision** which is generated by using a bilingual dictionary D to translate the right side of each training pair into the target language S to obtain $S' = \{(x_{e;i}, T_i, \mathbf{l}_i)\}$, where $T_i = \{t_{i1}, t_{i2}, .., t_{in}\}$ is the set of translations of $y_{e;i}$ in D. S' serves as weak supervision because semantic relations are not translation invariant (Figure 5.1), and hence the label $\mathbf{l}_i$ is not correct for every $(x_{e;i}, t_{ik})$ pair.

To extract useful training signals from the weak supervision, we use an **attention mechanism** which guides the model to attend to translations that preserve the monolingual label. The attention component constructs the input representation for the cross-lingual model M_{ef} in Equation 5.1 by averaging representations for all translation candidates, giving more weight to those that are likely to

preserve the monolingual label. Given a training sample $(x_{e;i},\ T_i,\ \mathbf{l}_i) \in S'$ with $T_i = \{t_{i1}, t_{i2}, .., t_{in}\}$, the score for a candidate translation t_{ik} is calculated using the word embeddings of $x_{e;i}$ and t_{ik}, along with $\mathbf{l}$, an embedding of the gold label $\mathbf{l}_i$ as features to a feed-forward network f (with one hidden layer). $\mathbf{l}$ is provided to help select translations that are consistent with the correct label for the monolingual pair. The scores for all translations are converted to probabilities using the softmax function, and the input features $\mathbf{v}'_{x_{e;i}y_{e;i}}$ for the student model M_{ef} are a sum of the features obtained from each of these translations, weighted by the probabilities.

$$score((x_{e;i}, t_{ik}), \mathbf{l}_i) = f([\mathbf{x}_i; \mathbf{t}_{ik}; \mathbf{l}]) \tag{5.3}$$

$$p(t_{ik}) = \frac{\exp(score(t_{ik}))}{\sum_{j=1}^{n} \exp(score(t_{ij}))}$$

$$\mathbf{v}'_{x_{e;i}y_{e;i}} = \sum_{k=1}^{n} p(t_{ik}) \mathbf{v}_{x_{e;i}t_{ik}} \tag{5.4}$$

The student model is then trained to maximize the **distillation objective**:

$$L_2 = -\sum_{j=1}^{c} [(1-\alpha)\ l_{ij}\ \log\ \hat{l}_j^{\,e\to f} + \alpha\ \hat{l}_j^{\,e\to e}\ \log(\frac{\hat{l}_j^{\,e\to f}}{\hat{l}_j^{\,e\to e}})] \tag{5.5}$$

where $\hat{\mathbf{l}}^{\,e\to e}$ is calculated using M_e, $\hat{\mathbf{l}}^{\,e\to f}$ is calculated using the attended representation $\mathbf{v}'_{x_{e;i}y_{e;i}}$ as input to M_{ef} and α is an interpolation parameter. The first term is again a cross-entropy loss that aims to measure how well the cross-lingual model M_{ef} does at predicting the relation given $\mathbf{v}'_{x_{e;i}y_{e;i}}$. The second term uses KL-Divergence (Kullback and Leibler, 1951) to penalize differences in predictions by M_{ef} on the cross-lingual input $\mathbf{v}'_{x_{e;i}y_{e;i}}$ and the predictions by M_e on the mono-

lingual input $\mathbf{v}_{x_{e;i}y_{e;i}}$. As is typical in distillation, we flatten the softmax of both inputs to the KL-Divergence term, by using a temperature parameter τ.

5.3 MultiLexRel : A Dataset for Cross-lingual Semantic Relations

Existing resources containing annotated cross-lingual lexical relations are limited in scope, quality, and quantity. Resources such as bilingual dictionaries or the Open Multilingual WordNet (Bond and Foster, 2013) can be mined for examples of synonyms, hypernyms and hyponyms, but these examples are noisy as these resources are created in a semi-automatic way.

In this section, we crowdsource MultiLexRel, a set of new high-quality annotations for English-Hindi (En-Hi) and English-Chinese (En-Zh) word pairs using the natural logic relations laid out in Section 6.2.[1] We leverage monolingual annotations to speed up the process and enable comparisons between monolingual and cross-lingual models. We use Google Translate to translate one side of a randomly sampled subset of the gold-standard dataset of semantic relations created by Pavlick et al. (2015), and ask crowdworkers whether the semantic relation holds after translation. Each example is annotated by five annotators and annotations are aggregated using MACE, a Bayesian model that estimates the trustworthiness of annotators and accordingly assigns a label to each instance (Hovy et al., 2013). The distributions of the five relations are shown in Table 5.1. 40-45% of the examples shown to annotators were deemed to not preserve the monolingual relation after translation. The final test sets consist of the remaining 55-60% examples.

[1] Via `http://figure-eight.com/`

Relation	En-Hi	En-Zh
`Equivalence`	158	174
`Forward Entail`	220	240
`Backward Entail`	215	236
`Exclusion`	124	154
`Other`	323	94
Total	1040	898

Table 5.1: Distribution of the five semantic relations for the two crowdsourced test sets.

5.4 Experimental Settings

MultiLexRel is used as a test set to evaluate our models. Training only requires English labeled examples, and other resources derived from raw monolingual and parallel corpora.

5.4.1 Data

English Supervision The English training samples are derived from the English Lexical-XXXL PPDB. After filtering away pairs containing non-alphabetic characters, we choose a random sample as training pairs. The number of samples for all classes is balanced, except for `Exclusion` (since there are fewer examples of this class in PPDB). All in all, the size of the training set is ∼20K pairs. Like previous work, we ensure a lexical split where the English words that are present in the test data are not seen in the training data (Levy et al., 2015). This makes the task challenging as it prevents model from memorizing patterns of words such as their "prototypicality" for certain relations *i.e.* whether certain words are likely to appear in specific relations.

Validation data Since we assume no access to labeled cross-lingual examples, we need to define a validation set using the resources available to us. We construct a validation set by randomly removing 1000 pairs from the training data, and automatically translating the right side of each example with the bilingual dictionary used for training. This process yields a noisy validation set, which is solely used for tuning hyper-parameters.

Unlabeled Resources The bilingual dictionary for knowledge distillation is obtained from the MUSE project (Lample et al., 2018) for En-Hi, while the MDBG dictionary is used for En-Zh.[2] We use FastText bilingual embeddings (Bojanowski et al., 2017).[3] We extract English paths for the monolingual model from the English Wikipedia.[4] Cross-lingual paths are extracted from a random sample of the WMT18 parallel corpora[5] for En-Zh (∼5M sentences) and the IIT Bombay English-Hindi corpus (Kunchukuttan et al., 2018) for En-Hi (∼1M sentences). All corpora are parsed using YaraParser (Rasooli and Tetreault, 2015) trained on the treebank of the corresponding language from the Universal Dependencies (v2.2) project (McDonald et al., 2013). Tokenization is performed using the Moses tokenizer for English (Koehn et al., 2007), the Indic NLP tokenizer for Hindi,[6] and the Jieba word segmenter for Chinese.[7]

[2] `https://www.mdbg.net/chinese/dictionary?page=cc-cedict`
[3] https://fasttext.cc/docs/en/aligned-vectors.html
[4] `https://dumps.wikimedia.org/enwiki/`
[5] `http://statmt.org/wmt18/translation-task.html`
[6] `https://github.com/anoopkunchukuttan/indic_nlp_library`
[7] `https://github.com/fxsjy/jieba`

5.4.2 Model Configurations and Baselines

Model Configuration The path-encoder LSTM has two layers with 60 hidden units each, with dropout (Srivastava et al., 2014) applied after the first layer. All feed-forward neural networks have a single hidden layer with 50 hidden units and dropout regularization. All models are trained in mini-batches of size 4 using the Adam optimizer (Kingma and Ba, 2014) with initial learning rate set to 10^{-3}. The temperature parameter τ for knowledge distillation is tuned over $\{1.0, 1.5, 2.0, 5.0\}$, and the interpolation parameter α over $\{0.75, 0.90\}$.

English-only Model: EnLexNet We also use a vanilla LexNet model applied to a monolingual test set in order to measure the gap between cross-lingual performance and monolingual performance. EnLexNet is trained on the same English samples used for training the BiLexNet model, and evaluated on the En-En examples used to generate cross-lingual examples in Section 5.3.We re-implement the LexNet model and verify its accuracy by replicating results on the CogALex dataset.

Baselines Our experiments aim to assess how the direct cross-lingual modeling of semantic relations in BiLexNet impacts predictions, and to isolate the impact of key training components: knowledge distillation and translation selection via attention. We compare against the following baselines :

- Random Baseline: Randomly assign one of the five semantic relations to each word pair.

- TRANSLATION BASELINE: This baseline combines dictionary translation and the English-only system ENLEXNET to gauge the relative difficulty of predicting semantic relations across languages rather than in English only. Each pair (x_e,y_f) in the test set is translated into English using the bilingual dictionary D. Since y_f can have multiple translations, we pair x_e with each of these translations, and use ENLEXNET to predict the relation for each of these pairs. The relation for (x_e, y_f) is then chosen as the most general relation among those predicted for the translated pairs according to the order in which they appear in Table 5.1. We also experimented with a voting based approach to combination, but it generally performed worse.

- BILEXNET (NO DISTILLATION): A simple strategy for cross-lingual transfer consists in seeding a vanilla LEXNET model with bilingual embeddings in the source and target languages before training. This strategy has been successfully used for other NLP tasks (Guo et al., 2015; Klementiev et al., 2012, *inter alia*). By keeping the embeddings fixed, we can use source language data to train the monolingual LEXNET model using features based on source embeddings and source language paths as usual. At inference time, the model uses both the source and target embeddings as input, and the cross-lingual paths defined above.

- SPECIALIZED TENSOR MODEL (STM): How does a model that has primarily been used for comparing words in the same language perform on cross-lingual comparisons? Our final baseline answers this question. Proposed by Glavaš

and Vulić (2018), STM is a neural architecture for identifying semantic relations that achieves state-of-the-art performance on two English datasets. STM is based on the hypothesis that specialized word embeddings are necessary to accurately disambiguate between semantic relations.

More precisely, STM assumes that different specializations of generic word embeddings are needed to recognize different relations and that interactions between the specialized vectors can be used to identify the semantic relations. These different specializations are implemented using K feed-forward neural networks. Given a word pair, STM takes in as input a pair of generic word embeddings for the word pair which are then specialized by the K transformations. Each pair of corresponding specialized embeddings is used to calculate a score based on a non-linear transformation of their bilinear product. Finally, the K scores obtained from K pairs of specialized embeddings are used as features to train a multi-class classifier.

Besides English, STM has also been used for cross-lingual *transfer*, where a model trained on one language (say English) is used to test on word-pairs in another language (say German). Here, we use STS in a new setting to predict the semantic relation between two words in different languages.

We use the official implementation of STM with the same bilingual embeddings used by BiLexNet.[8] We tune three hyperparameters on the validation set: the size of the specialized tensors {100, 200, 300, 500}, the number of spe-

[8] `https://github.com/codogogo/stm`

Model	En-Hi			En-Zh		
	P	R	F	P	R	F
Random Baseline	22.9±1.3	20.9±2.5	21.4±1.4	22.9±1.3	20.9±2.5	21.4±1.4
Translation Baseline	30.1±1.7	26.3±1.3	28.3±1.5	50.1±2.2	32.8±0.8	33.0±1.0
STM	32.0	33.0	29.0	20.0	15.0	16.0
BiLexNet (No distillation)	34.9±1.2	34.3±0.6	32.2±0.8	41.7±6.4	33.5±6.2	32.6±7.2
BiLexNet (No attention)	47.2±1.0	42.9±1.8	41.9±2.3	45.0±1.5	40.8±1.8	39.8±0.8
BiLexNet (Full)	47.7±1.2	44.2±0.8	**43.3±1.0**	48.3±1.6	41.6±1.1	**41.1±0.9**

Table 5.2: Precision (P), Recall (R) and F1-score (F) for BiLexNet and contrastive baselines on the two MultiLexRel test sets. All configurations are trained with five random seeds. We report the mean score and standard deviation. The full BiLexNet model performs best and is consistently better with the attention component.

cialization functions {3, 5, 7}, and the learning rate {0.0001, 0.0003}. Default values are used for all other hyper-parameters.

5.5 Results

Tables 5.2 summarizes results on the MultiLexRel test sets. BiLexNet achieves F1 scores that are roughly double of those obtained by the random baseline for 5-way classification.

Impact of cross-lingual modeling We assess the impact of direct cross-lingual modeling in BiLexNet by comparing against the Translation Baseline. Using a translation dictionary to naïvely convert cross-lingual relation prediction to an English task, the translation baseline F1 scores are 8 to 13 points higher than Random for both language pairs. This difference can be attributed to easy examples where English semantic relations are preserved by simple dictionary translation. BiLexNet further improves F1 by 8 to 15 points over the Translation Base-

LINE, primarily by improving recall.

Supervised English system Without cross-lingual training samples, we cannot compare weakly supervised and fully supervised training for BILEXNET in a controlled fashion. However, the supervised monolingual ENLEXNET model (Section 5.4) evaluated on the En-En test set offers a reference point: remarkably the F1 scores of BILEXNET are only 1 to 3 points lower than those obtained by the supervised English model (∼44 on the En-En test set).

Impact of knowledge distillation We compare the full BILEXNET model to the naïve baseline (BILEXNET (NO DISTILLATION)) that only relies on embeddings for cross-lingual transfer and does not perform cross-lingual distillation. This approach performs on par with or a little better than the translation baseline, but ∼9 points worse than the full BILEXNET model, losing on both precision and recall. This result confirms the benefit of aligning training and test conditions for our model with knowledge distillation and not relying solely on embeddings. These results are consistent with prior findings on distributional representations.

- Distributional representations have difficulties in discriminating between multiple semantic relations (Chersoni et al., 2016a). As such, relying solely on word embeddings for cross-lingual transfer can cause loss of knowledge during transfer.

- Syntactic divergences cause differences in paths in the source and target languages. This can cause a distribution shift between the features seen by the

classifier during training and test time, thereby affecting performance. Again, word embeddings are not sufficient to bridge the gap between the distributions of the two languages (Chen et al., 2018b).

Impact of attention We test the impact of the attention model in BiLexNet by removing it, and instead translating training samples for distillation using the single most frequent translation. Removing attention yields small but consistent degradations, suggesting that attending to multiple translations is beneficial, but leaves room for improvement. We analyze the behavior of the attention model in the next section.

Specialization Finally, we observe F1 scores of STM are significantly worse than those of BiLexNet. In fact, it is the weakest model for En-Zh, and is only 3 points better than the translation baseline for En-Hi. The relatively poor performance of STM highlights that our cross-lingual task, which directly compares words in two languages, is fundamentally different from the transfer task, where models trained in one language are ported to other languages. Thus, models such as STM, which have been designed for transfer, may not be directly suitable for our task.

5.6 Analysis

This section further breaks down the results, and highlight some successes and failures of BiLexNet to guide future work on cross-lingual semantic relation classification.

Class	En-Hi	En-Zh	En-En
`Equivalence`	33	30	31
`Exclusion`	33	28	23
`Forward Entail`	47	48	48
`Backward Entail`	45	58	48
`Other`	51	29	53

Table 5.3: Per-class F1 scores for median En-Hi and En-Zh BiLexNet model and the EnLexNET model.

Performance Per Class We break down the performance of the BiLexNet model per target relation (Table 5.3). The `Equivalence` and `Exclusion` classes are the hardest to predict correctly, which is consistent with our monolingual results and those from prior work (Shwartz and Dagan, 2016c): distributional models have trouble distinguishing synonyms from antonyms (Yih et al., 2012) and synonyms rarely occur in the same sentence, and hence path-based methods are less useful for this class. However, in BiLexNet, words of the `Equivalence` class can occur in a parallel sentence pair where they are aligned to each other. Thus, there is a direct signal for examples of this class which helps discriminate between `Equivalence` and `Exclusion`.

The largest fraction of errors are caused by the model predicting `Other` instead of a specific relation. This suggests that special treatment of this class might improve performance, perhaps by using a multi-step process which filters out pairs not related under the relations that we are targeting, and then performs 4-way classification for the remaining examples. This is similar to the the CogALex shared task, where the first part of the task is to eliminate *completely unrelated* pairs, before predicting relations on the remaining pairs (Santus et al., 2016). However, filtering

out *unrelated* pairs is an easier task than filtering pairs in the `Other` category.

Missing cross-lingual paths Cross-lingual paths might not exist for all word pairs, particularly for language pairs with limited parallel data such as En-Hi. BiLexNet would then only rely on word embeddings as features to predict semantic relations. We assess the impact of missing paths by comparing the classification performance on pairs which have cross-lingual paths (70% of the test), against pairs which do not have paths in the En-Hi setting. The former subset has a higher F1 score (44.6) than the latter (40.2), mainly due to differences in recall. This difference in performance also confirms that the cross-lingual paths complement word embeddings, in the same way that monolingual paths do.

Attention Analysis We complement ablation experiments in Table 5.2 by examining a random sample of 25 monolingual training pairs (x_e, y_e) where y_e has multiple translations in the bilingual dictionary. We manually check for how many pairs the model places the highest attention weight on a translation that preserves the relation label of the monolingual pair. This happens in 64% of the cases (16 out of 25). The attention model is often able to modulate the choice of the right translation of y_e based on the context provided by x_e and the gold label. For example, given the monolingual example (*drop*, *fall*, `Forward Entail`) the model places the highest weight on the Hindi word गिरा, which captures the "moving downward" sense. On the other hand, for the example (*autumn*, *fall*, `Equivalence`), the model correctly identifies पतझड as the right translation.

There still remains a lot of overhead for improving the attention component. Some failure cases in the 25 examples occur for pairs where the set of translations of y_e contains an incorrect translation which is totally unrelated to x_e or y_e. For example, given (*country*, *uganda*), the model chooses the word कैंडल (transliteration for *candle*) and not युगांडा (transliteration for *uganda*). Of course, this is an extreme example, but such errors are also more likely to occur when the noisy translation is in the same domain as x_e and y_e. Fixing such errors can help improve the training process.

5.7 A Comparison of Unsupervised and Weakly Supervised Approaches

Having studied the performance of BiLexNet on the task of identifying the full spectrum of semantic relations from Section 6.2, we now compare its performance against the BiSparse-Dep model introduced in Chapter 4, on the task of cross-lingual hypernymy detection. Besides the empirical question of which method is more accurate at identifying hypernyms, there are two other reasons why such a comparison is useful.

First, since both models assume access to different kinds of resources, this comparison can guide model choice based on the availability of resources. Specifically, recall that BiSparse-Dep requires a small amount of cross-lingual word pairs labeled with hypernymy information for tuning hyper-parameters, which are potentially difficult to create for new language pairs. On the other hand, BiLexNet uses a large number of English training pairs labeled with semantic relations, from an al-

	Hyper-Hypo		Hyper-Cohypo	
	En-Fr	En-Zh	En-Fr	En-Zh
BiSparse (Unsupervised)	59.9	57.0	66.5	67.3
BiLexNet (Weakly Supervised)	70.9±0.81	70.7±1.17	69.0±2.79	73.5±4.9

Table 5.4: Accuracies of the weakly supervised BiLexNet when compared to the unsupervised BiSparse method from Chapter 4 on the task of cross-lingual hypernymy detection. The weak supervison enables BiLexNet to more accurately identify cross-lingual hypernymy than the unsupervised approach.

ready existing resource *viz.* PPDB. Conversely, BiLexNet needs sentence-aligned parallel data, while BiSparse-Dep can be trained using a bilingual dictionary. Knowing the trade-off between resources and accuracy can help in better selecting which approach to use for new language pairs.

Second, work in monolingual settings has established that unsupervised methods for hypernymy detection are generally less accurate than their supervised counterparts (Levy et al., 2015; Shwartz et al., 2017). Comparing the two approaches will reveal whether and to what extent this is true in cross-lingual settings.

Setup We use the English-Chinese and English-French subsets of the Hyper-Cohypo and Hyper-Hypo test beds described in Section 4.4. Our training setup for MultiLexRel is exactly the same as described earlier (Section 5.4), with one small change. We match the binary classification setup of the cross-lingual hypernymy detection task by treating the training and validation examples with `Forward Entail` label as positive examples, and all other examples as negative examples.

Results Table 5.4 shows that BiLexNet is more accurate than BiSparse-Dep for both languages tested and on both versions of the dataset. Improvements are

larger on the HYPER-HYPO datasets, but performances are similar across HYPER-HYPO and HYPER-COHYPO. This indicates that BILEXNET is able to distinguish hypernyms from hyponyms as well as hypernymy from co-hyponymys with the same ease. In contrast, BISPARSE-DEP is more accurate at distinguishing hypernyms from co-hyponyms than hyponymys (Tables 4.3 and 4.4).

These results also confirm that the merits of supervision for hypernymy detection hold even in cross-lingual settings. Despite being only weakly supervised using English examples translated via dictionary-guided distillation, BILEXNET is more accurate than the unsupervised BISPARSE-DEP.

5.8 Summary

This chapter motivated and introduced the task of classifying semantic relations between words in different languages with the objective of precisely characterizing differences in word meaning. Additionally, we introduced MULTILEXREL, a dataset of about 1000 English-Hindi and 900 English-Chinese word pairs annotated with the natural logic lexical entailment classes of MacCartney and Manning (2007), and BILEXNET, a cross-lingual relation classification model.

We also introduced a knowledge distillation algorithm for BILEXNET, which only needs annotated monolingual examples and a bilingual dictionary. Unlike previous uses of knowledge distillation for cross-lingual transfer, our approach does not assume that labels are translation invariant, and relies on an attention mechanism to select translations that best explain a given label. Experiments show that this

method largely outperforms baselines that use bilingual embeddings or dictionaries more naïvely for cross-lingual transfer, and that it approaches the performance of fully supervised systems on an English-only version of the task.

Taken together, the last two chapters provide two different ways of characterizing differences in meaning between words in two different languages. While we discuss one potential use-case of modeling semantic relations in a later chapter (Chapter 7), such models can be used for a wide variety of problems in multilingual NLP by discovering fine-grained characterization of difference in meaning across languages.

Chapter 6: Identifying Semantic Divergences in Parallel Text without Annotations

Having introduced models to detect semantic relations between words across languages, we now turn to the problem of identifying differences in meaning in bilingual sentence pairs drawn from parallel corpora. In this chapter, we discuss an automatic method to distinguish semantically equivalent sentence pairs from semantically divergent pairs, so that parallel corpora can be used more judiciously in downstream cross-lingual NLP applications. We propose a semantic model to automatically detect whether a sentence pair is semantically divergent (Section 6.2). While prior work relies on surface cues to detect misaligments, our approach focuses on comparing the meaning of words and overlapping text spans using bilingual word embeddings (Luong et al., 2015) and a deep convolutional neural network (He and Lin, 2016). Crucially, training this model requires no manual annotation. Noisy supervision is obtained automatically borrowing techniques developed for parallel sentence extraction (Munteanu and Marcu, 2005). Our model can thus easily be trained to detect semantic divergences in any parallel corpus.

We extensively evaluate our semantically-motivated models on the intrinsic task of detecting divergent examples in the two parallel English-French data sets we

collected in Chapter 3 (Section 3.2). We show that our semantically motivated model significantly outperforms other methods based on word alignment cues, shallow word embeddings, and neural machine translation scores.

6.1 Background

Non-Parallel Corpora Beyond understanding the impact of noisy data (Section 2.1.4), mismatches in bilingual sentence pairs have also been studied to extract parallel segments from non-parallel corpora to augment MT training data (Abdul-Rauf and Schwenk, 2009; Fung and Cheung, 2004; Munteanu and Marcu, 2005, 2006; Riesa and Marcu, 2012; Smith et al., 2010, *inter alia)*. Methods for parallel sentence extraction rely primarily on surface features based on translation lexicons and word alignment patterns (Munteanu and Marcu, 2005, 2006). Similar features have proved to be useful for the related task of translation **quality estimation** (Specia et al., 2010, 2018), which aims to detect divergences introduced by MT errors, rather than human translation. Recently, sentence embeddings have also been used to detect parallelism (España-Bonet et al., 2017; Schwenk and Douze, 2017). Although embeddings capture semantic generalizations, these models are trained with neural MT objectives, which do not distinguish semantically equivalent segments from divergent parallel segments.

Cross-lingual Semantics Tasks such as cross-lingual semantic textual similarity (STS) (Agirre et al., 2014) and cross-lingual textual entailment (CLTE) (Mehdad et al., 2010; Negri et al., 2012, 2013) seek to characterize semantic relations between

sentences in two different languages beyond translation equivalence, and are therefore directly relevant to our goal of identifying divergences in bilingual sentences. We re-purpose data and methods for CLTE to determine whether annotations for closely related cross-lingual tasks help in identifying divergences (Section 6.3.5). Our core model for detecting divergences, which we introduce in the next section, is based on an accurate model for identifying monolingual semantic similarity.

6.2 Approach

We introduce our approach to detecting divergence in parallel sentences, with the goal of (1) detecting differences ranging from large mismatches to subtle nuances, (2) without manual annotation.

Cross-Lingual Semantic Similarity Model We address the first requirement using a neural model that compares the meaning of sentences using a range of granularities. We re-purpose the Very Deep Pairwise Interaction (VDPWI) model, which has been previously been used to detect semantic textual similarity (STS) between English sentence pairs (He and Lin, 2016). It achieved competitive performance on data from the STS 2014 shared task (Agirre et al., 2014), and outperformed previous approaches on sentence classification tasks (He et al., 2015; Tai et al., 2015) with fewer parameters, faster training, and without requiring expensive external resources such as WordNet.

The VDPWI model was designed for comparing the meaning of sentences in the same language, based not only on word-to-word similarity comparisons, but

also on comparisons between overlapping spans of the two sentences, as learned by a deep convolutional neural network. We adapt the model to our cross-lingual task by initializing it with bilingual embeddings. To the best of our knowledge, this is the first time this model has been used for cross-lingual tasks in such a way. We give a brief overview of the resulting model here and refer the reader to the original paper for details. Given sentences e and f, VDPWI models the semantic similarity between them using a pipeline consisting of five components:

1. **Bilingual word embeddings**: Each word in e and f is represented as a vector using pre-trained, bilingual embeddings.

2. **BiLSTM for contextualizing words**: Contextualized representations for words in e and f are obtained by choosing the output vectors at each time step obtained by running a bidirectional LSTM (Schuster and Paliwal, 1997) on each sentence.

3. **Word similarity cube**: The contextualized representations are used to calculate various similarity scores between each word in e with each word in f. Each score thus forms a matrix and all such matrices are stacked to form a *similarity cube* tensor.

4. **Similarity focus layer**: The similarity cube is fed to a similarity focus layer that re-weights the similarities in the cube to focus on highly similar word pairs, by decreasing the weights of pairs which are less similar. This output is called the *focus cube*.

5. **Deep convolutional network**: The focus cube is treated as an "image" and passed to a deep neural network, the likes of which have been used to detect patterns in images. The network consists of repeating convolution and pooling layers. Each repetition consists of a spatial convolutional layer, a Rectified Linear Unit (Nair and Hinton, 2010), and a max pooling layer, followed by fully connected layers, and a softmax to obtain the final output.

The entire architecture is trained end-to-end to minimize the Kullback-Leibler divergence (Kullback and Leibler, 1951) between the output similarity score and gold similarity score.

Noisy Synthetic Supervision How can we obtain gold similarity scores as supervision for our task? We automatically construct examples of semantically divergent and equivalent sentences as follows. Since a large number of parallel sentence pairs are semantically equivalent, we use parallel sentences as positive examples. Synthetic negative examples are generated following the protocol introduced by Munteanu and Marcu (2005) to distinguish parallel from non-parallel segments. Specifically, candidate negative examples are generated starting from the positive examples $\{(e_i, f_i)\ \forall i\}$ and taking the Cartesian product of the two sides of the positive examples $\{(e_i, f_j)\ \forall i, j \text{ s.t. } i \neq j\}$. This candidate set is filtered to ensure that negative examples are not too easy to identify: we only retain pairs that are close to each other in length (a length ratio of at most 1:2), and have enough words (at least half) which have a translation in the other sentence according to a bilingual dictionary derived from automatic word alignments.

This process yields positive and negative examples that are a noisy source of supervision for our task, as some of the positive examples might not be fully equivalent in meaning. However, experiments in the next section (Section 6.4) will show that, in aggregate, they provide a useful signal for the VDPWI model to learn to detect semantic distinctions.

6.3 Divergence Detection Evaluation

We evaluate the accuracy of our cross-lingual semantic divergence detector using the dataset annotated in Chapter 3 (Section 3.2), and compare it against a diverse set of baselines in controlled settings. We test our hypothesis that semantic divergences are more than alignment mismatches by comparing the semantic divergence detector with models that capture mis-alignment (Section 6.3.2) or translation (Section 6.3.3). Then, we compare the deep convolutional architecture of the semantic divergence model, with a simpler model that directly compares bilingual sentence embeddings (Section 6.3.4). Finally, we compare our model trained on synthetic examples with a supervised classifier used in prior work to predict finer-grained textual entailment categories based on manually created training examples (Section 6.3.5). Except for the entailment classifier which uses external resources, all models are trained on the exact same parallel corpora (OpenSubtitles or CommonCrawl for evaluating on the corresponding test bed). This is a transductive setup, where the test set is a (very small) subset of the training set. Note that the true label for a sentence pair, *i.e.* whether it is divergent or not, is not known during training.

6.3.1 Neural Semantic Divergence Detection

Model and Training Settings We use the publicly available implementation of the VDPWI model.[1] We initialize with 200 dimensional Bivec French-English word embeddings (Luong et al., 2015), trained on the parallel corpus from which our test set is drawn. We use the default setting for all other VDPWI parameters. The model is trained for 25 epochs and the model that achieves the best Pearson correlation coefficient on the validation set is chosen. At test time, VDPWI outputs a score $\in [0, 1]$, where a higher value indicates less divergence. We tune a threshold on development data to convert the real-valued score to binary predictions.

Synthetic Data Generation The synthetic training data is constructed using a random sample of 5000 sentences from the training parallel corpus as positive examples. We generate negative examples automatically as described in Section 6.2, and sample a subset to maintain a 1:5 ratio of positive to negative examples. We experimented with other ratios and found that the results only slightly degraded while using a more balanced ratio (1:1, 1:2), but severely degraded with a skewed ratio (1:9).

6.3.2 Parallel vs. Non-parallel Classifier

Are divergences observed in parallel corpora more than alignment errors? To answer this question, we reimplement the model proposed by Munteanu and Marcu

[1] `https://github.com/castorini/VDPWI-NN-Torch`

(2005). It discriminates parallel pairs from non-parallel pairs in comparable corpora using a supervised linear classifier with the following features for each sentence pair (e, f):

- Length features: $|f|$, $|e|$,$\frac{|f|}{|e|}$, and $\frac{|e|}{|f|}$
- Alignment features (for each of e and f): Alignments are obtained using IBM Model 2 trained in each direction, combined with `union`, `intersection`, and `grow-diag-final-and` heuristics, and the following features are extracted:
  - Count and ratio of unaligned words
  - Top three largest fertilities
  - Longest contiguous unaligned and aligned sequence lengths
- Dictionary features: A bilingual dictionary is constructed using word alignments from a random sample of a million sentences from the training parallel corpus. Features measure fraction of words in e that have a translation in f and vice-versa.

Training uses the exact same synthetic examples as the semantic divergence model (Section 6.2).

6.3.3 Neural MT

If divergent examples are nothing more than bad translations, a neural MT system should assign lower scores to divergent segments pairs than to those that are equivalent in meaning. We test this empirically using neural MT systems trained

for a single epoch, and use the system to score each of the sentence pairs in the test sets. We tune a threshold on the development set to convert scores to binary predictions.

The MT system uses the attentional encoder-decoder model (Bahdanau et al., 2014) implemented in the Sockeye toolkit (Hieber et al., 2017). Encoders and decoders are single-layer GRUs with 1000 hidden units (Cho et al., 2014). Source and target word embeddings have size 512. Using byte-pair encoding, the vocabulary size is 50000 (Sennrich et al., 2016). Maximum sequence length is set to 50.

We optimize the standard cross-entropy loss using Adam (Kingma and Ba, 2014), with learning rate set to 0.0003 and halved when the validation perplexity does not decrease for 3 checkpoints. The batch size is set to 80. Preliminary experiments showed that training for more than one epoch does not help divergence detection, so we terminate training after an epoch.

6.3.4 Bilingual Sentence Embeddings

Our semantic divergence model introduces a large number of parameters to combine the pairwise word comparisons into a single sentence-level prediction. This baseline tests whether a simpler model would suffice. We detect semantic divergence by computing the cosine similarity between sentence embeddings in a bilingual space. The sentence embeddings are bag-of-word representations, built by taking the mean of bilingual word embeddings for each word in the sentence. This approach has been shown to be effective, despite ignoring fundamental linguistic information such as

word order and syntax (Mitchell and Lapata, 2010). We use the same 200 dimensional Bivec word embeddings (Luong et al., 2015), trained on OpenSubtitles and CommonCrawl respectively.

6.3.5 Textual Entailment Classifier

Our final baseline re-purposes annotations and models designed for the task of Cross-Lingual Textual Entailment (CLTE) to detect semantic divergences. This baseline was introduced in Carpuat et al. (2017), and it helps us understand how the synthetic training data compares to training examples generated manually, for a related cross-lingual task. Using CLTE datasets from SemEval (Negri et al., 2012, 2013), we train a supervised linear classifier that can distinguish sentence pairs that entail each other, from pairs where entailment does not hold in at least one direction. The features of the classifier are based on word alignments and sentence lengths.

First, differences in sentence lengths are strong indicators of divergence between e and f. Accordingly, we use four length features: $|e|$, $|f|$, $\frac{|e|}{|f|}$, and $\frac{|f|}{|e|}$.

Second, we assume that the configuration of word alignment links between parallel sentences (e, f) is indicative of equivalence: if e and f have the same meaning, then they will be easier to align. Accordingly, we compute the following features for each of e and f:

- Ratio of aligned words
- Ratio of unaligned words
- Ratio of unaligned content words (defined as words that do not appear in a

stopword list)

- Number of unaligned contiguous sequences
- Length of longest contiguous unaligned sequence
- Average length of aligned sequences
- Average length of unaligned sequences

All alignments are trained on 2M sentence pairs from Europarl (Koehn, 2005) using the Berkeley aligner (DeNero and Klein, 2007; Liang et al., 2006). The classifier is the linear SVM from Scikit-Learn.[2]

6.4 Intrinsic Evaluation Results

Table 6.1 shows that the semantic similarity model is most successful at distinguishing equivalent from divergent examples. The break down per class shows that both equivalent and divergent examples are better detected. The improvement is larger for divergent examples with gains of about 10 points for F-score for the divergent class, when compared to the next-best scores.

Among the baseline methods, the non-entailment model is the weakest. While it uses manually annotated training examples, these examples are drawn from distant domains, and the categories do not exactly match the task at hand. In contrast, the other models benefit from training on examples drawn from the same corpus as each test set.

Next, the MT based model and the sentence embedding model, both of which

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html`

Divergence Detection Approach	+P	+R	+F	-P	-R	-F	Overall F
OpenSubtitles							
Sentence Embeddings	65	60	62	56	61	58	60
MT Scores (1 epoch)	67	53	59	54	68	60	60
Non-entailment	58	78	66	53	30	38	54
Non-parallel	70	83	76	61	42	50	66
Semantic Dissimilarity	**76**	**80**	**78**	**75**	**70**	**72**	**77**
Common Crawl							
Sentence Embeddings	78	58	66	52	**74**	61	64
MT Scores (1 epoch)	54	65	59	17	11	14	42
Non-entailment	73	49	58	48	72	57	58
Non-parallel	70	83	76	61	42	49	67
Semantic Dissimilarity	**82**	**88**	**85**	**78**	69	**73**	**80**

Table 6.1: Intrinsic evaluation on crowdsourced semantic equivalence vs. divergence testsets. We report overall F-score, as well as precision (P), recall (R) and F-score (F) for the equivalent (+) and divergent (-) classes separately. Semantic similarity more accurately identifies divergences across the board, with larger improvements on the divergent class.

are unsupervised, are significantly weaker than the two supervised models trained on synthetic data, highlighting the benefits of the automatically constructed divergence examples. The strength of the semantic similarity model compared to the sentence embeddings model highlights the benefits of the fine-grained representation of bilingual sentence pairs as a similarity cube, as opposed to the bag-of-words sentence embedding representation.

Finally, despite training on the same noisy synthetic data as the parallel vs non-parallel system, the semantic similarity model is better able to detect divergences. This highlights the benefits of (1) meaning comparison between words in a shared embedding space, over the discrete translation dictionary used by the baseline, and (2) the deep convolutional neural network which enables the comparison of overlapping spans in sentence pairs, as opposed to local word alignment features.

6.4.1 Analysis

We manually examine the 13-15% of examples in each test set that are correctly detected as divergent by semantic similarity and mis-classified by the non-parallel detector.

On OpenSubtitles, most of these examples are true divergences rather than noisy alignments (*i.e.* sentences that are not translation of each other). The non-parallel detector weighs length features highly, and is fooled by sentence pairs of similar length that share little content and therefore have very sparse word alignments. The remaining sentence pairs are plausible translations in some context that still contain inherent divergences, such as details missing or added in one language. The non-parallel detector views these pairs as non-divergent since most words can be aligned. The semantic similarity model can identify subtle meaning differences, and correctly classifies them as divergent. As a result, the non-parallel detector has a higher false positive rate (22%) than the semantic similarity classifier (14%), while having similar false negative rates (11% v/s 12%).

On the CommonCrawl test set, the examples with disagreement are more diverse, ranging from noisy segments that should not be aligned to sentences with subtle divergences.

6.5 Summary

This chapter focused on semantic divergences in parallel sentences and introduced an approach for detecting such divergences. The model is based on neural

semantic similarity and importantly, does not require manual annotation, and thus can be trained for any language pair and domain with a parallel corpus. Evaluating the model on the intrinsic divergence datasets collected in Chapter 3 shows that the model detects such divergences much more accurately than shallower translation or alignment based models. While divergence detection accuracy can still be further improved, a relevant questions is how does the intrinsic performance translate to improvements on a downstream task? We answer this question in the next chapter, where we investigate the impact of divergence detection on machine translation.

Chapter 7: Divergences in NLP Applications

Over the last three chapters, we have introduced models for identifying lexico-semantic relations and semantic divergences across languages. Further, these models have been tested intrinsically for their ability to identify the phenomena they are meant to capture. In this chapter, we take an extrinsic view and study semantic divergences in the context of downstream tasks.

In the first part of this chapter, we consider semantic divergences between words in the context of the task of bilingual dictionary induction. Using this task allows us to identify whether predictions from BiLexNet can distinguish semantically equivalent translations from divergent translations in settings where bilingual word pairs are relatively closer in meaning than in MultiLexRel, thus exhibiting more subtle divergences.

In the latter part, we focus on semantic divergences between sentences. We use our model for detecting semantic divergences in parallel data (Section 6.2) to identify divergences in data used to train a neural MT system. We use the model for data selection for English-French and Vietnamese-English machine translation (MT), and show that the models helps select data to train neural MT models faster with no loss in translation quality.

Taken together, this chapter provides evidence that semantic divergences affect downstream NLP tasks and motivates building improved models for detecting such divergences.

7.1 Semantic Divergences in Automatically Constructed Dictionaries

The testbed used to evaluate BiLexNet (Section 5.3) consists of arbitrary word pairs that are not restricted to be close in meaning. However, semantic relations also exist between word pairs that are closer in meaning than those in MultiLexRel. For instance, near-synonyms, which are very similar but not identical in meaning, also exhibit relations such as `Forward Entail`/`Backward Entail` (execute-kill) and `Exclusion` (execute-murder) (Edmonds and Hirst, 2002). Cruse (1986) terms relations between near-synonyms as *micro-relations*, as the intensity of these relations is diminished when compared to word-pairs that are not near-synonyms. Thus, micro-relations represent a more subtler notion of divergences than macro-relations (which are captured in MultiLexRel).

Experiments in Section 5.5 on MultiLexRel demonstrate that BiLexNet identifies macro-relations more accurately than naive baselines based on translation or bilingual embeddings. Can BiLexNet also identify micro-relations? In this section, we use the task of *bilingual dictionary induction* (BDI) to answer this question. BDI consists of building a lexicon of bilingual word pairs that are equivalent in meaning. Thus, word pairs drawn from automatically constructed lexicons using BDI are expected to be close in meaning. At the same time, as observed by Peirsman

and Padó (2011), such word pairs also exhibit a variety of semantic relations. This tension makes the task a suitable candidate for testing the ability of BiLexNet to identify micro-relations. Specifically, we ask whether predictions from BiLexNet can improve the accuracy of BDI, by separating semantically divergent translations, represented by micro-relations, from equivalent translations.

7.1.1 Background: Bilingual Dictionary Induction via Bilingual Embeddings

Formally, BDI is defined over a source language E (with vocabulary V_e) and a target language F (with vocabulary V_f). Given a set of words in the source language, $S_e = \{e_1, e_2, ..., e_k\}$, the task is to find words $S_f = \{f_1, f_2, ..., f_k\} \subset V_f$ such that $\forall i$, e_i and f_i have the same meaning. BDI has been widely studied, under a variety of assumptions about the resources used (Irvine and Callison-Burch, 2017; Klementiev et al., 2012; Rapp, 1995, *inter alia)*. Recent work has motivated BDI as an evaluation task for bilingual embeddings, and this has in-turn led to purely embeddings-based approaches to this task. We operate under such an embeddings-based framework since it is conceptually simple and better highlights the utility of modeling semantic relations to capture divergences. Methods based on embeddings are also known to outperform more traditional approaches such as those based on multilingual topic models (Vulić and Moens, 2016).

The typical approach to BDI using word embeddings consists of two steps:

1. **Construction of a bilingual word embeddings space**: Each word w in

both V_e and V_f is represented using a vector $\vec{w}$ in a common vector space.

2. **Identifying translations**: For each $e_i \in S_e$, a translation $f_i \in V_f$ is identified using the positions of vectors of words in V_f with respect to the position of $\vec{e_i}$. One way to do this is by identifying the target language word f, such that $\vec{f}$ is most similar to $\vec{e_i}$ in the embedding space *i.e.* $f_i = \arg\max_{f \in V_f} \vec{e_i} \cdot \vec{f}$. Other methods such as the inverted softmax (Smith et al., 2017) and cross-domain similarity local scaling (CSLS) (Conneau et al., 2017) have been proposed as alternative distance metrics but we focus on cosine similarity for simplicity.

7.1.2 Using BiLexNet to Filter Divergent Lexical Translations

We propose a modification to the embeddings-based approach described above that combines evidence from word embeddings and the output of BiLexNet. The objective, as stated earlier, is to separate equivalent translations from non-equivalent translations, and thereby test if BiLexNet can capture micro-relations. However, results in Section 5.5 show that BiLexNet does not always generate correct predictions. To extract meaningful signal from the noisy predictions of BiLexNet, we make the following two modeling assumptions:

1. Instead of using the predicted relations directly, we use the full range of output probabilities over the five semantic relations, with the expectation that these probabilities contain more information than the raw predictions.

2. While similarity is not the same as equivalence, we recognize that the cosine similarity between word embeddings is often a strong indicator of semantic

equivalence. Hence, instead on solely relying on the BiLexNet probabilities, we use an evidence combination approach that combines these probabilities with cosine similarities between embeddings.

Formally, given a query word, e_i, our approach consists of the following steps

1. **Generating Candidate Sets** : Obtain the K most-similar words to e in the target language. Let this set be $C_f = \{f_{i1}, f_{i2}, ..., f_{iK}\} \subset V_f$, and the associated similarities be $Sim_f = \{sim_{i1}, sim_{i2}, ..., sim_{iK}\}$. C_f is our candidate set—we assume that the right translation exists in this set, and we aim to identify it by re-ranking the candidates in the next steps.[1]

2. **Scoring Candidates using BiLexNet**: Using BiLexNet, we calculate $syn(e_i, f_{ik})\ \forall k \in [1, K]$, *i.e.* the probability that the candidate is equivalent in meaning to the query word. $syn(.)$ directly aims to model equivalence in meaning in contrast to cosine similarity.

3. **Normalizing Scores**: We normalize both the similarity scores and the BiLexNet probabilities using the softmax function to obtain two independent probability distributions over the candidate set. Thus,

$$p_{sim}(e_i, f_{ik}) = \frac{\exp(sim_{ik})}{\sum_{k'=1}^{K} \exp(sim_{ik'})} \quad \text{and} \quad p_{syn}(e_i, f_{ik}) = \frac{\exp(syn(e, f_{ik})}{\sum_{k'=1}^{K} \exp(syn(e, f_{ik'}))} \tag{7.1}$$

4. **Evidence Combination**: We combine these two independent sources of in-

[1] K is set to 10 for all experiments.

Approach	Full Test Set		Manually Filtered Test Set	
	Precision@1	Precision@5	Precision@1	Precision@5
EMBEDDINGS	45.9	63.4	51.7	70.3
BILEXNET	27.8	50.1	30.3	57.0
EMBEDDINGS + BILEXNET	**47.9**	63.4	**53.1**	70.8
ORACLE (Precision@10)	67.1		74.2	

Table 7.1: Combining predictions of semantic relations from BILEXNET with embeddings-based cosine similarity consistently improves Precision@1 for bilingual dictionary induction. Results in **bold** are statistically significant compared to next best result in the column (McNemar's test, $p < 0.05$).

formation to assign a final score for each candidate. Thus

$$score(e_i, f_{ik}) = p_{sim}(e_i, f_{ik}) + p_{syn}(e_i, f_{ik}) \tag{7.2}$$

The correct translation is the candidate one with the highest score *i.e.*

$$f_i^* = \arg\max_{f_i \in C_f} score(e_i, f_{ik}) \tag{7.3}$$

7.1.3 Setup and Data

For evaluation, we use the Hindi-English bilingual dictionary from (Pavlick et al., 2014) that we studied in Section 3.1. The full dictionary consists of 9150 Hindi words, paired with one or more English translations. We filter the dictionary to keep only lexical translation into English (While multi-word translations are important they are not within the ambit of this work). We also filter out stop words in both languages, words containing numerals, and source words that contain characters not in the Devanagari script. Our final test set consists of 7764 Hindi

words. Besides evaluating on this complete set, we also evaluate on the subset that was manually labeled as being equivalent in Section 3.1. This subset provides a more cleaner testbed, as we saw in Section 3.1 that an unfiltered dictionary is likely to contain divergences, which can make the evaluation unreliable. To obtain C_f, we use the MUSE embeddings as we did in the intrinsic experiments in Section 5.4. Similarly, we obtain BiLexNet predictions from the model trained as described in Section 5.4.

We contrast results obtained from the approach described in Section 7.1.2 (Embeddings + BiLexNet) with results from the two components that we combine:

1. Embeddings: The most-similar translation from C_f
2. BiLexNet: The translation with the highest $syn(*)$ score

As is standard, we measure performance on the task using Precision@1 and Precision@5. Additionally, we also calculate an upper-bound on any re-ranking based approach by measuring the Oracle precision, which measures how many query words have a right translation in C_f.

7.1.4 Results and Discussion

Comparing the Oracle approach (P@1 = 67.1) with the Embeddings only approach (P@1 = 46.9) shows that a large absolute gain of almost 20 points is possible with a re-ranking approach. Unsurprisingly, the BiLexNet translations by themselves, with a P@1 of 30.5, are the least accurate amongst the methods

considered. This highlights the need for better models for semantic relations that can simultaneously perform well on intrinsic and extrinsic evaluations.

Most importantly, combining EMBEDDINGS and BILEXNET leads to a small but statistically significant boost (McNemar's test, $p < 0.05$) over EMBEDDINGS. A cleaner test set boosts all scores by about 3 to 7 points, as demonstrated by the scores on the smaller, manually curated set. However, even with higher scores, the difference in performance between EMBEDDINGS and EMBEDDINGS + BILEXNET remains constant (and significant).

Taken together, these two sets of results demonstrate that despite not being fully accurate, predictions from BILEXNET provide useful signals that capture micro-relations and filter out divergent relations. These predictions refine an embeddings-based approach and more accurately detect semantically equivalent relationships across languages.

7.2 Improving Neural MT Training by Filtering Semantic Divergences

Having seen the impact of semantic divergences model on a word-level downstream task, we now turn our attention to sentence-level divergences, and study the utility of our divergence detector (Section 6.2) on a downstream task. Specifically, we focus on machine translation. We study neural machine translation since, as discussed in Section 2.1.4, neural MT systems are more sensitive to the nature of examples than phrase based-systems. Concretely, we hypothesize that divergent

training pairs can hamper a neural MT system and that filtering such pairs can help in better training of neural MT systems. We investigate this hypothesis by taking a data selection approach, and selecting the least divergent examples in a parallel corpus based on a range of divergence detectors discussed in Section 6.3, and comparing the translation quality of the resulting neural MT systems.

7.2.1 Translation Tasks

English-French We evaluate on 4867 sentences from the Microsoft Spoken Language Translation dataset (Federmann and Lewis, 2016) as well as on 1300 sentences from TED talks (Cettolo et al., 2012). Training examples are drawn from OpenSubtitles, which contains ∼28M examples after deduplication. We discard 50% examples for data selection.

Vietnamese-English Since the SEMANTIC SIMILARITY model was designed to be easily portable to new language pairs, we also test its impact on the IWSLT Vietnamese-English TED task, which comes with ∼120,000 and 1268 in-domain sentences for training and testing respectively (Cettolo et al., 2016). This is a more challenging translation task as Vietnamese and English are more distant languages, there is little training data, and the sentence pairs are expected to be cleaner and more parallel than those from OpenSubtitles. In these lower resource settings, we discard 10% of examples for data selection.

7.2.2 Neural MT System

The MT system is the same one used in the intrinsic experiments for the neural MT based divergence detector (Section 6.3.3). Neural architectures such as the one we use here are significantly stronger than phrase-based systems on high-resource language pairs such as English-French (Cettolo et al., 2016) and are also competitive in low-resource settings (Luong and Manning, 2015).[2] We use the attentional encoder-decoder model (Bahdanau et al., 2014) implemented in the Sockeye toolkit (Hieber et al., 2017). Encoders and decoders are single-layer GRUs (Cho et al., 2014) with 1000 hidden units. Source and target word embeddings have size 512. Using byte-pair encoding (Sennrich et al., 2016), the vocabulary size is 50000. Maximum sequence length is set to 50.

We optimize the standard cross-entropy loss using Adam (Kingma and Ba, 2014), until validation perplexity does not decrease for 8 checkpoints. The learning rate is set to 0.0003 and is halved when the validation perplexity does not decrease for 3 checkpoints. The batch size is set to 80. At decoding time, we construct a new model by averaging the parameters for the 8 checkpoints with best validation perplexity, and decode with a beam of 5. All experiments are run thrice with distinct random seeds.

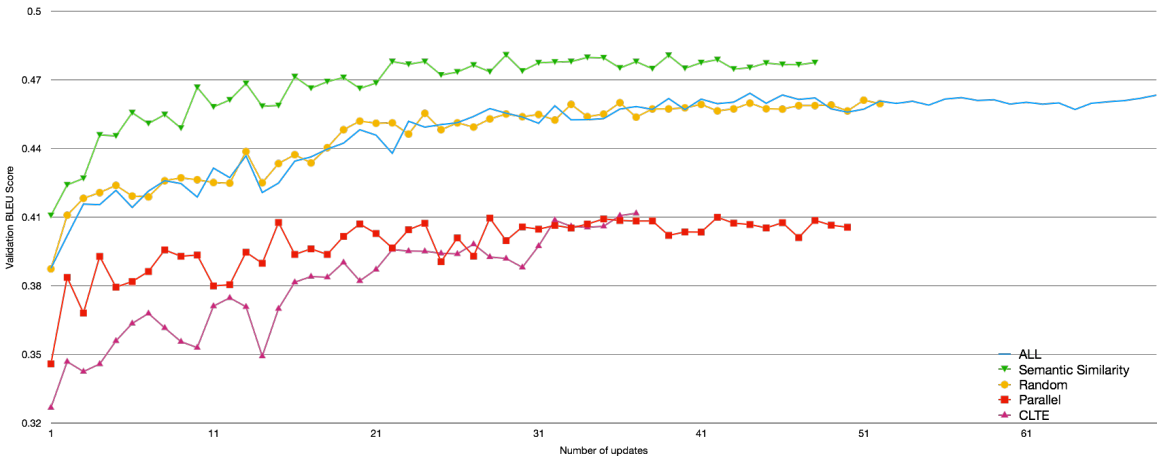

Figure 7.1: Learning curves on the validation set for English-French models (mean of 3 runs/model). The SEMANTIC SIMILARITY model outperforms other models throughout training, including the one trained on all data.

Model	**MSLT BLEU**		**TED BLEU**	
	Average	Ensemble	Average	Ensemble
RANDOM	43.49	45.64	36.05	38.20
PARALLEL	40.65	42.12	35.99	37.86
ENTAILMENT	39.64	41.86	33.30	35.40
SEMANTIC SIMILARITY	**45.53**	**47.23***	**36.98**	**38.87**
ALL	44.64	46.26	36.98	38.59

Table 7.2: English-French decoding results. BLEU scores are either averaged across 3 runs ("Avgerage") or obtained via ensemble decoding ("Ensemble"). SEMANTIC SIMILARITY reach BLEU scores on par with ALL with only half of the training data. SEMANTIC SIMILARITY scores marked with * are significanly better ($p < 0.05$) than the corresponding ALL scores.

7.2.3 English-French Results

We train English-French neural MT systems by selecting the least divergent half of the training corpus with the following criteria:

- SEMANTIC SIMILARITY (Section 6.2)
- PARALLEL: the non-parallel sentence detector (Section 6.3.2)

[2]More recent work (Sennrich and Zhang, 2019) has shown that a carefully tuned neural system can outperform phrase-based systems even in low-resource settings.

- ENTAILMENT: the entailment classifier (Section 6.3.5)

Learning curves (Figure 7.1) show that data selected using SEMANTIC SIMILARITY yields better validation BLEU throughout training compared to all other models. SEMANTIC SIMILARITY selects more useful examples for MT than PARALLEL, even though both selection models are trained on the same synthetic examples. This highlights the benefits of semantic modeling over surface mis-alignment features. Furthermore, SEMANTIC SIMILARITY achieves the final validation BLEU of the model that uses ALL data with only 30% of the updates. This suggests that semantically divergent examples are pervasive in the training corpus, confirming the findings from manual annotation (Section 3.2), and that the presence of such examples slows down neural MT training.

Decoding results on the blind test sets (Table 7.2) show that SEMANTIC SIMILARITY outperforms all other data selection criteria (with differences being statistically significant under bootstrap resampling tests (Koehn, 2004), $p < 0.05$), and performs as well or better than the ALL model which has access to twice as many training examples.

7.2.4 Vietnamese-English Results

Trends from English-French carry over to Vietnamese English, as the SEMANTIC SIMILARITY model compares favorably to ALL while reducing the number of training updates by 10%. SEMANTIC SIMILARITY also yields better BLEU than RANDOM with the differences being statistically significant. While differences in

Model	Avgerage Test Set BLEU
Random (90%)	22.71
Semantic Sim. (90%)	**23.38**
All	23.30

Table 7.3: Vietnamese-English decoding results: dropping 10% of the data based on Semantic Similarity does not hurt BLEU, and performs significantly ($p < 0.05$) better than Random selection.

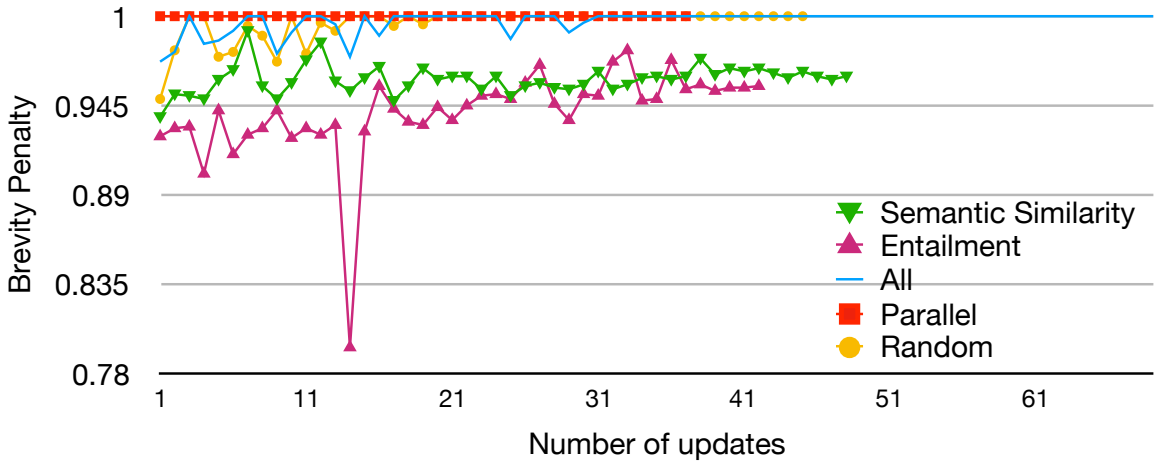

Figure 7.2: Brevity penalties on the validation set for English-French models.

score here are smaller, these results are encouraging since they demonstrate that our semantic divergence models port to more distant low-resource language pairs.

7.2.5 Analysis

We break down the results seen in Figure 7.1 and Table 7.2, with a focus on the behavior of the Entailment and All models. We start by analyzing the BLEU brevity penalty trends observed on the validation set during training (Figure 7.2).

We observe that both the Entailment and Semantic Similarity based models have similar brevity penalties despite having performances that are at opposite ends of the spectrum in terms of BLEU. This implies that translations generated

Entailment is inadequate due to under-translation	
Source	he's a very impressive man **and still goes out to do digs.**
Reference	c'est un homme très impressionnant **et il fait encore des fouilles.**
ENTAILMENT	c'est un homme très impressionnant.
Source	when the Heat **first** won.
Reference	lorsque les Heat ont gagné **pour la première fois**.
ENTAILMENT	quand le Heat a gagné.
Parallel produces garbage tokens	
Source	alright.
Reference	d'accord.
ENTAILMENT	**{ \ pos (192,210)}** d'accord.

Table 7.4: Selected translation examples from the ensemble systems of the various models.

by the SEMANTIC SIMILARITY model have better n-gram overlap with the reference, but are much shorter. Manual examination of the translations suggests that the ENTAILMENT model often fails by under-translating sentences, either dropping segments from the beginning or the end of source sentences (Table 7.4).

The PARALLEL model consistently produces translations that are longer than the reference. This is partially due to the model's propensity to generate a sequence of garbage tokens in the beginning of a sentence, especially while translating shorter sentences. In our test set, almost 12% of the translated sentences were found to begin with the garbage text shown in Table 7.4. Only a small fraction ($< 0.02\%$) of the French sentences in our training data begin with these tokens, but the tendency of PARALLEL to promote divergent examples above non-divergent ones, seems to exaggerate the generation of this sequence.

7.3 Summary

In this chapter, we moved beyond intrinsic evaluation for models of semantic divergences, and investigated how semantic divergences and our models for detecting differences in meaning across languages impact downstream tasks. First, using the task of bilingual dictionary induction, we investigated the ability of BiLexNet to recognize micro-relations and showed that predictions of BiLexNet can be combined with cosine similarity scores to improve the accuracy of an automatically induced dictionary by filtering out divergent translations. Second, we showed that divergences in parallel data impede the training of neural MT systems, and that filtering out divergences reduces the training time of such systems by half, without affecting the translation quality. Collectively, results from this chapter demonstrate that semantic divergences manifest in various ways in downstream tasks, and affect the task in different ways. However, automatic models that do not rely on task-specific labeled data can help in alleviating the impact of such divergences.

Chapter 8: Conclusion

The main claim of this thesis is that translations are not always meaning preserving, and as a result cross-lingual semantic divergences are pervasive in multilingual settings. Translation ambiguity prevents direct porting of monolingual models and datasets to cross-lingual settings, and hence this thesis contributes novel cross-lingual tasks, datasets, and models which directly identify and characterize differences in meaning between two words and sentences in different languages.

We conclude by summarizing the central contributions of this thesis and pointing out key limitations of the work discussed here which suggest directions for future research.

8.1 Summary of Contributions

We start in Chapter 3 by showing that translations recorded in bilingual resources are not always meaning preserving, leading to semantic divergences. Annotating subsets of data from a bilingual dictionary and two parallel corpora reveals that divergences discovered cover a wide spectrum, including subtle differences in meaning, well-defined taxonomic relations, as well as noisy translations. Parts of this chapter were previously published in Vyas et al. (2018).

Motivated by the fact that translation is not meaning preserving, we introduce novel tasks to directly identify and characterize differences in meaning across languages. In Chapters 4 and 5, we focused on word level differences in meaning and proposed two novel tasks with the objective of exactly characterizing such differences by identifying the semantic relation between two words. The first task (Chapter 4) is that of identifying cross-lingual hypernymy, as hypernymy is commonly observed in automatically induced translation pairs (Peirsman and Padó, 2011), while the second task expands to simultaneously classifying between multiple lexico-semantic relations from the natural logic framework (MacCartney and Manning, 2007, 2009) (Chapter 5). These tasks have been well-studied in purely monolingual settings, but datasets and models cannot directly port to cross-lingual setting because changes in meaning due to translations can change semantic relations. Instead, we directly annotate bilingual word pairs for these relations, and create datasets spanning multiple languages for both tasks (Sections 4.4 and 5.3).

Solving these cross-lingual tasks by converting them to monolingual tasks in a high resource language via translation is complicated by the exact same reason—translation does not preserve meaning and this can cause semantic relations to change. Thus, we introduce solutions for these tasks do not rely on direct translation, and instead encode the ambiguity of translation into the models. Our models also eschew traditional supervised learning as labeled cross-lingual data for these tasks is difficult to obtain. Instead, we recognize hypernymy through an unsupervised model of sparse, bilingual embeddings that encodes the distributional inclusion hypothesis (Geffet and Dagan, 2005). Our model, BiSparse-Dep, learns from large

monolingual corpora, and aligns the two languages using a bilingual dictionary that allows for multiple translations for a word (Section 4.3). For the broader task of identifying between multiple lexico-semantic relations, we present BiLexNet, a neural model that predicts relations by combining bilingual word embeddings, with lexico-syntactic paths in both languages (Section 5.1). Unlike BiSparse-Dep, which was an unsupervised model, BiLexNet is trained in a weakly supervised fashion using labeled examples in English which are easily available. These examples are used to train a cross-lingual predictor using a novel dictionary-guided knowledge distillation approach (Section 5.2). Chapters 4 contains work that was previously published in Vyas and Carpuat (2016) and Upadhyay et al. (2018),[1] while work in Chapter 5 appears in Vyas and Carpuat (2019).

We shifted our attention to semantic divergences in parallel sentences in Chapter 6 and introduced an approach for directly identifying if a sentence and its translation are semantically equivalent. The model, which is based on neural semantic similarity, does not require manual annotation and thus can be trained for any language pair and domain with a parallel corpus. Evaluating the model on the intrinsic divergence datasets collected in Chapter 3 shows that our model detects such divergences much more accurately than shallower translation or alignment based models. This work previously appeared in Vyas et al. (2018).

Finally, in Chapter 7 we showed that semantic divergences affect downstream tasks, but models developed in this thesis alleviate the impact of such divergences. We first modeled word-level divergences in the task of bilingual dictionary induction

[1]The first two authors contributed equally to this work

using BiLexNet, and showed that even when the input word pair is close in meaning, BiLexNet helps separate divergent translations from equivalent translations. We also showed that divergences in parallel sentences slow down training of neural MT systems, and that filtering out divergences reduces the training time of such systems by half, without affecting the translation quality. Portions of this chapter were previously published in Vyas et al. (2018).

8.2 Limitations and Future Work

In this section, we point out some limitations of the work discussed in this thesis. These limitations naturally raise some open questions which can serve as directions for future work.

8.2.1 Fine-grained Distinction of Divergences

A limitation of our work on detecting divergences in parallel data, is that we do not distinguish between different *types* of divergences *i.e.* we only consider whether a sentence pair as either divergent or not. This is a narrow view of divergences because translators adapt different translation strategies, and these different strategies can give rise to different types of divergences (Baker, 2011). Treating different types of divergences in the same way may not be an optimal way of dealing with these differences. The work of Zhai et al. (2018, 2019) is a step in this direction, as they investigate how different translation processes are manifested in real parallel data, and to what extent can they be recognized automatically. However, their approach

relies on labeled training data, which they point out is very expensive and time-consuming to obtain. Whether such fine-grained information can be detected using weaker forms of supervision (*e.g.* synthetic data as in our work), or even in an unsupervised way, is a crucial open question for expanding to more language pairs.

Distinguishing types of divergences is also crucial for downstream tasks, as this information can guide how different divergences in data are handled. For example, while we show that discarding the most divergent half of a parallel corpus improves training time of a neural MT system, this is a crude strategy that discards ∼14 million parallel sentences, including some which are potentially valuable. Knowing why a divergent pair is divergent can help in retaining useful pairs and potentially extract training signals from them. This strategy has been adapted by Pham et al. (2018), who build upon our work and show that certain semantic divergences, *viz.* those that contain extra words on either the source or the target side, can be fixed (by deleting those extra words), and these fixed examples can be used for training instead of discarding. Knowing the type of semantic divergence can also help in defining curriculum learning strategies (Zhang et al., 2018, 2019), where training examples are presented in a well-defined order based on their divergence type and score. Such strategies could lead to more judicious use of training data, which is especially important in low-resource settings where parallel data of any form is hard to acquire.

Similarly, while studying lexico-semantic relations, we only make categorical distinctions, *e.g.* either a word is a hypernym of another word or it is not. However, work on semantic prototypes has established certain concepts are consistently more

central to a specific semantic category than others (Rosch, 1975), and thus semantic relations can be captured on a graded scale. This graded notion has been explored for monolingual hypernymy detection (Vulić et al., 2017) where high inter-annotator agreement scores reveal that even non-experts are consistent in identifying the degree of hypernymy. Vulić et al. (2019) also extend our work to build models for graded cross-lingual hypernymy, but it remains to be seen how other cross-lingual semantic relations can be identified in a graded fashion and how such graded judgments can assist downstream applications.

8.2.2 Integrating Lexical and Sentential Models by Identifying Semantic Relations in Context

Approaches for identifying semantic relations between words and semantic divergences between sentences are discussed independently in this thesis, but the problems themselves are not. As a simple example, knowing that the Spanish *lavar* (wash) is a hypernym of the English *shampoo*, allows us to identify that the two sentences "*lavar el cabello*" (wash the hair) and "*shampoo the hair*" are semantically divergent. This raises the question of whether models of lexical relations offer ways of improving sentence level models by identifying fine-grained information between words in sentences. A related challenge is the ability to model lexical semantics in context. We investigate cross-lingual semantic relations in Chapters 4 and 5 at the word type level, without taking into account the context in which these words appear. This obstructs the integration of word and sentence level models because

lexico-semantic relations can change with context, based on the sense in which the word was used.

Admittedly, identifying semantic relations in context is a challenging problem even within the same language, and has been relatively less explored in prior work compared to the context-agnostic tasks. In our own work, we have shown that existing methods for monolingual, context-agnostic hypernymy detection, when used in tandem with contextualized word representations, perform only slightly better than context-agnostic baselines (Vyas and Carpuat, 2017). This motivates the need for specialized methods for the context-sensitive version of the task. Deep contextualized word representations obtained using large-scale, language-modeling based pre-training can potentially serve as strong baselines for such context-sensitive tasks(Devlin et al., 2019; Peters et al., 2018). For the broader task of identifying semantic relation in context, Shwartz and Dagan (2016a) introduced the CONTEXT-PPDB dataset. This dataset consists of word pairs along with a pair of sentential contexts, with a label indicating the semantic relation between the two words in the given contexts (the labels are those used in PPDB (Pavlick et al., 2015)). However, since CONTEXT-PPDB only consists of ∼3700 sentence pairs, it provides only a smaller number of annotated examples per relation, making it difficult to train large supervised models on. Such challenges of models and data are likely to be exacerbated in multilingual and cross-lingual settings.

A related limitation of the work on lexico-semantic relations, both in this thesis as well as in the broader literature in monolingual settings, is that models for representing words using vectors typically conflate different senses of a word

into a single representation. Sense-agnostic word representations cannot capture all senses of a word even when such senses occur in the corpus used to train these embeddings (Schütze, 1998; Yaghoobzadeh and Schütze, 2016), and this can cause inaccurate detection of semantic relations which are typically defined between senses. Possible solutions to this problem include modeling word meaning using multi-sense representations that more accurately capture multiple senses of a single word type (Camacho-Collados and Pilehvar, 2018), or by explicitly balancing the various senses while building representations (Yin and Roth, 2018).

8.2.3 Impact of Divergences and Cross-lingual Semantic Relations on other Cross-lingual Tasks

A final question of interest is to understand the impact of semantic divergences and cross-lingual semantic relations on tasks beyond translation.

Semantic relations have been found to be useful in many downstream monolingual tasks. They have been used for relating named entities in question answering (McNamee et al., 2008), query expansion in information retrieval (Voorhees, 1994), and for tasks such as coreference resolution (Ponzetto and Strube, 2006) and relation extraction (Demeester et al., 2016). Similarly, can relations between words in different languages assist cross-lingual versions of these tasks (*e.g.* cross-lingual information retrieval, or cross-lingual question answering)?

Bilingual dictionaries and parallel corpora are used to facilitate cross-lingual learning by transferring labeled data (Hwa et al., 2005; Mayhew et al., 2017; Yarowsky

et al., 2001) and trained models (Kozhevnikov and Titov, 2013; McDonald et al., 2011) from one language to another for many NLP tasks such as part-of-speech tagging, named entity recognition, semantic role labeling, and syntactic parsing. To what extent do semantic divergences affect cross-lingual transfer of both models as well as data? How can we systematically measure the impact of divergences on a diverse array of tasks? Can we build robust multilingual models that ameliorate the impact of these divergences? Answering these and other related questions will help in understanding the broader impact of divergences on multilingual NLP problems.

Bibliography

Sadaf Abdul-Rauf and Holger Schwenk. On the Use of Comparable Corpora to Improve SMT Performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 16–23, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

Željko Agić and Natalie Schluter. Baselines and Test Data for Cross-Lingual Inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May 2018. European Languages Resources Association (ELRA).

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 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, August 2014. Association for Computational Linguistics and Dublin City University.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 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, 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1081.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. Many Languages, One Parser. *Transactions of the Association for Computational Linguistics*, 4:431–444, December 2016. ISSN 2307-387X. doi: 10.1162/tacl_a_00109.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Mona Baker. *In Other Words : A Coursebook on Translation*. Routledge, January 2011. ISBN 978-1-136-83973-3. doi: 10.4324/9780203832929.

Colin Bannard and Chris Callison-Burch. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 597–604, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219914.

James Barnett, Inderjeet Mani, Elaine Rich, Chinatsu Aone, Kevin Knight, and Juan Carlos Martinez. Capturing language-specific semantic distinctions in interlingua-based MT. In *In Proceedings, Machine Translation*, pages 25–32, 1991.

Marco Baroni and Alessandro Lenci. Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):673–721, December 2010. ISSN 0891-2017, 1530-9312. doi: 10.1162/coli_a_00016.

Marco Baroni and Alessandro Lenci. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10, Edinburgh, UK, July 2011. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, September 2009. ISSN 1574-0218. doi: 10.1007/s10579-009-9081-4.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32, Avignon, France, April 2012. Association for Computational Linguistics.

Robert De Beaugrande and Wolfgang U. Dressler. *Introduction to Text Linguistics*. Longman, 1981. ISBN 978-0-582-55486-3.

Yonatan Belinkov and Yonatan Bisk. Synthetic and Natural Noise Both Break Neural Machine Translation. In *arXiv:1711.02173 [Cs]*, November 2017.

Luisa Bentivogli and Emanuele Pianta. Looking for lexical gaps. In *Proceedings of the 9th EURALEX International Congress*, pages 663–669, 2000. ISBN 978-3-00-006574-3.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus Phrase-Based Machine Translation Quality: A Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1025.

Matthew Berland and Eugene Charniak. Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 57–64. Association for Computational Linguistics, 1999.

Rahul Bhagat and Eduard Hovy. What Is a Paraphrase? *Computational Linguistics*, 39(3):463–472, May 2013. ISSN 0891-2017. doi: 10.1162/COLI_a_00166.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, December 2017. ISSN 2307-387X. doi: 10.1162/tacl_a_00051.

Francis Bond and Ryan Foster. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. Adding dense, weighted connections to wordnet. In *In Proceedings of the Third International WordNet Conference*, 2006.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, pages 169–176, Stroudsburg, PA, USA, 1991. Association for Computational Linguistics. doi: 10.3115/981344.981366.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993.

John A. Bullinaria and Joseph P. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526, August 2007. ISSN 1554-3528. doi: 10.3758/BF03193020.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June 2012. Association for Computational Linguistics.

Jose Camacho-Collados and Mohammad Taher Pilehvar. From Word To Sense Embeddings: A Survey on Vector Representations of Meaning. *Journal of Artificial Intelligence Research*, 63:743–788, December 2018. ISSN 1076-9757. doi: 10.1613/jair.1.11259.

Sharon A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 120–126. Association for Computational Linguistics, 1999.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. Detecting Cross-Lingual Semantic Divergence for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, August 2017. doi: 10.18653/v1/W17-3209.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 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, 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001.

M Cettolo, J Niehues, S Stuker, L Bentivogli, R Cattoni, and M Federico. The IWSLT 2016 Evaluation Campaign. page 14, 2016.

Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th EAMT Conference*, pages 261–268, Trento, Italy, 2012.

Boxing Chen, Roland Kuhn, George Foster, Colin Cherry, and Fei Huang. Bilingual Methods for Adaptive Training Data Selection for Machine Translation. volume Proc. of AMTA, pages 99–103, 2016.

Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. XL-NBT: A Cross-lingual Neural Belief Tracking Framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 414–424, 2018a. doi: 10.18653/v1/D18-1038.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification. *Transactions of the Association for Computational Linguistics*, 6:557–570, December 2018b. ISSN 2307-387X. doi: 10.1162/tacl_a_00039.

Emmanuele Chersoni, Giulia Rambelli, and Enrico Santus. CogALex-V Shared Task: ROOT18. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 98–103, Osaka, Japan, December 2016a. The COLING 2016 Organizing Committee.

Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. Representing Verbs with Rich Contexts: An Evaluation on Verb Similarity. In *Empirical Methods in Natural Language Processing*, pages 1967–1972. Association for Computational Linguistics, 2016b. doi: 10.18653/v1/D16-1205.

Andrew Chesterman. *Memes of Translation: The Spread of Ideas in Translation Theory*. J. Benjamins, June 1997. ISBN 978-1-55619-706-2.

David Chiang. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228, June 2007. ISSN 0891-2017, 1530-9312. doi: 10.1162/coli.2007.33.2.201.

Timothy Chklovski and Patrick Pantel. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of EMNLP 2014*, June 2014.

Jinho D. Choi, Joel Tetreault, and Amanda Stent. It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool. 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 387–396, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1038.

Kenneth W. Church, William Gale, Patrick Hanks, Donald Hindle, and Rosamund Moon. *Lexical Substitutability*. 1994.

Eve V. Clark. The principle of contrast: A constraint on language acquisition. In *Mechanisms of Language Aquisition*, pages 1–33. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1987. ISBN 978-0-89859-596-3 978-0-89859-973-2.

Daoud Clarke. Context-theoretic Semantics for Natural Language: An Overview. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 112–119, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word Translation Without Parallel Data. *arXiv:1710.04087 [cs]*, October 2017.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269.

Courtney Corley and Rada Mihalcea. Measuring the Semantic Similarity of Texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, EMSEE '05, pages 13–18, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

D. A. Cruse. *Lexical Semantics*. Cambridge University Press, 1986.

Ido Dagan and Oren Glickman. PROBABILISTIC TEXTUAL ENTAILMENT: GENERIC APPLIED MODELING OF LANGUAGE VARIABILITY. page 6, 2004.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220, July 2013. ISSN 1947-4040. doi: 10.2200/S00509ED1V01Y201305HLT023.

Shachi Dave, Jignashu Parikh, and Pushpak Bhattacharyya. Interlingua-based English–Hindi Machine Translation and Language Divergence. *Machine Translation*, 16(4):251–304, December 2001. ISSN 1573-0573. doi: 10.1023/A:1021902704523.

Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. Lifted Rule Injection for Relation Embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1389–1399, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1146.

John DeNero and Dan Klein. Tailoring Word Alignments to Syntactic Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 17–24, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Dun Deng and Nianwen Xue. Translation Divergences in Chinese-english Machine Translation: An Empirical Investigation. *Comput. Linguist.*, 43(3):521–565, September 2017. ISSN 0891-2017. doi: 10.1162/COLI_a_00292.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. doi: 10.3115/1220355.1220406.

Bonnie Dorr. SOLVING THEMATIC DIVERGENCES IN MACHINE TRANSLATION. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 127–134, June 1990. doi: 10.3115/981823.981840.

Bonnie Dorr. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics*, 20(4):597–633, 1994.

Loïc Dugast, Jean Senellart, and Philipp Koehn. Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07*, pages 220–223, Prague, Czech Republic, 2007. Association for Computational Linguistics. doi: 10.3115/1626355.1626387.

Philip Edmonds and Graeme Hirst. Near-Synonymy and Lexical Choice. *Computational Linguistics*, 28(2):105–144, June 2002. ISSN 0891-2017. doi: 10.1162/089120102760173625.

C. España-Bonet, Á C. Varga, A. Barrón-Cedeño, and J. van Genabith. An Empirical Analysis of NMT-Derived Interlingual Embeddings and Their Use in Parallel Sentence Identification. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1340–1350, December 2017. ISSN 1932-4553. doi: 10.1109/JSTSP.2017.2764273.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. Sparse Overcomplete Word Vector Representations. 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 1491–1500, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1144.

Christian Federmann and William D Lewis. Microsoft Speech Language Translation (MSLT) Corpus: The IWSLT 2016 release for English, French and German. In *International Workshop on Spoken Language Translation (IWSLT)*, page 6, 2016.

Christiane Fellbaum. WordNet. In Roberto Poli, Michael Healy, and Achilles Kameas, editors, *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer Netherlands, Dordrecht, 2010. ISBN 978-90-481-8847-5. doi: 10.1007/978-90-481-8847-5_10.

John Rupert Firth. *A Synopsis of Linguistic Theory, 1930-1955*. 1957.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. ISSN 1939-1455(Electronic),0033-2909(Print). doi: 10.1037/h0031619.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning Semantic Hierarchies via Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1199–1209, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Pascale Fung and Percy Cheung. 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, August 2004. COLING.

William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics* -, pages 177–184, Berkeley, California, 1991. Association for Computational Linguistics. doi: 10.3115/981344.981367.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The Paraphrase Database. In *HLT-NAACL*, pages 758–764, 2013.

Luca Gasparri and Diego Marconi. Word Meaning. https://plato.stanford.edu/archives/fall2019/entries/word-meaning/#Bib, June 2015.

Maayan Geffet and Ido Dagan. Feature vector quality and distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics - COLING '04*, pages 247–es, Geneva, Switzerland, 2004. Association for Computational Linguistics. doi: 10.3115/1220355.1220391.

Maayan Geffet and Ido Dagan. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 107–114. Association for Computational Linguistics, 2005.

Goran Glavaš and Simone Paolo Ponzetto. Dual Tensor Model for Detecting Asymmetric Lexico-Semantic Relations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1757–1767, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. Discriminating between Lexico-Semantic Relations with the Specialization Tensor Model. 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 181–187, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2029.

Goran Glavaš and Ivan Vulić. Generalized Tuning of Distributional Word Vectors for Monolingual and Cross-Lingual Lexical Entailment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4824–4830, Florence, Italy, July 2019. Association for Computational Linguistics.

Tom Goldstein, Christoph Studer, and Richard Baraniuk. A field guide to forward-backward splitting with a FASTA implementation. *arXiv preprint arXiv:1411.3406*, 2014.

G. Golub and W. Kahan. Calculating the Singular Values and Pseudo-Inverse of a Matrix. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, 2(2):205–224, January 1965. ISSN 0887-459X. doi: 10.1137/0702016.

Cyril Goutte, Marine Carpuat, and George Foster. The Impact of Sentence Alignment Errors on Phrase-Based Machine Translation Performance. In *Proceedings of AMTA-2012: The Tenth Biennial Conference of the Association for Machine Translation in the Americas*, page 9, 2012.

Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA, 1994. ISBN 978-0-7923-9468-6.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. Cross-lingual Dependency Parsing Based on Distributed Representations. 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 1234–1244. Association for Computational Linguistics, 2015. doi: 10.3115/v1/P15-1119.

Zellig S. Harris. Distributional Structure. *WORD*, 10(2-3):146–162, August 1954. ISSN 0043-7956. doi: 10.1080/00437956.1954.11659520.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv:1803.05567 [cs]*, March 2018.

Hua He and Jimmy Lin. Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 937–948, San Diego, California, June 2016. Association for Computational Linguistics.

Hua He, Kevin Gimpel, and Jimmy Lin. 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, September 2015. Association for Computational Linguistics.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992.

Karl Moritz Hermann and Phil Blunsom. Multilingual Distributed Representations without Word Alignment. *arXiv:1312.6173 [cs]*, December 2013.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. Sockeye: A Toolkit for Neural Machine Translation. *arXiv:1712.05690 [cs, stat]*, December 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.

Juliane House. *Translation Quality Assessment: A Model Revisited*. Gunter Narr. Verlag, 1997. ISBN 978-3-8233-5075-0.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning Whom to Trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130. Association for Computational Linguistics, 2013.

Zhiheng Huang, Marcus Thint, and Zengchang Qin. Question Classification using Head Words and their Hypernyms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 927–936, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325, September 2005. ISSN 1469-8110, 1351-3249. doi: 10.1017/S1351324905003840.

Ann Irvine and Chris Callison-Burch. A Comprehensive Analysis of Bilingual Lexicon Induction. *Computational Linguistics*, 43(2):273–310, June 2017. ISSN 0891-2017, 1530-9312. doi: 10.1162/COLI_a_00284.

Sergio Jimenez, George Dueñas, Julia Baquero, and Alexander Gelbukh. UNAL-NLP: Combining Soft Cardinality Features for Semantic Textual Similarity, Relatedness and Entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 732–742, Dublin, Ireland, 2014. Association for Computational Linguistics. doi: 10.3115/v1/S14-2131.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *arXiv:1412.6980 [Cs]*, December 2014.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing Crosslingual Distributed Representations of Words. In *Proceedings of COLING 2012*, pages 1459–1474. The COLING 2012 Organizing Committee, 2012.

Philipp Koehn. Statistical Significance Tests for Machine Translation Evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of The Tenth Machine Translation Summit*, page 8, Phuket, Thailand, 2005.

Philipp Koehn and Rebecca Knowles. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3204.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 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. Association for Computational Linguistics, 2007.

Philipp Koehn, Kenneth Heafield, Mikel L. Forcada, Miquel Esplà-Gomis, Sergio Ortiz-Rojas, Gema Ramírez Sánchez, Víctor M. Sánchez Cartagena, Barry Haddow, Marta Bañón, Marek Střelec, Anna Samiotou, and Amir Kamran. ParaCrawl Corpus version 1.0. *http://paracrawl.eu*, January 2018a.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 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, 2018b. Association for Computational Linguistics. doi: 10.18653/v1/W18-6453.

Werner Koller. *Einführung in die Übersetzungswissenschaft*. Quelle & Meyer, Heidelberg, 1979. ISBN 978-3-494-02089-1. OCLC: 611363864.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. Directional distributional similarity for lexical expansion. page 69. Association for Computational Linguistics, 2009. doi: 10.3115/1667583.1667606.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(04):359–389, October 2010. ISSN 1351-3249, 1469-8110. doi: 10.1017/S1351324910000124.

Mikhail Kozhevnikov and Ivan Titov. Cross-lingual Transfer of Semantic Role Labeling Models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1200, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177729694.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The IIT Bombay English-Hindi Parallel Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May 2018. European Languages Resources Association (ELRA).

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. February 2018.

Els Lefever, Marjan Van de Kauter, and Véronique Hoste. Evaluation of Automatic Hypernym Extraction from Technical Corpora in English and Dutch. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 490–497, Reykjavik, Iceland, May 2014. European Languages Resources Association (ELRA).

Alessandro Lenci and Giulia Benotto. Identifying Hypernyms in Distributional Semantic Spaces. In *Proceedings of 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 '12, pages 75–79, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. Do Supervised Distributional Methods Really Learn Lexical Inference Relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado, 2015. Association for Computational Linguistics.

Percy Liang, Ben Taskar, and Dan Klein. Alignment by agreement. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* -, pages 104–111, New York, New York, 2006. Association for Computational Linguistics. doi: 10.3115/1220835.1220849.

Dekang Lin. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 768–774, Montreal, Quebec, Canada, August 1998. Association for Computational Linguistics. doi: 10.3115/980691.980696.

Pierre Lison and Jorg Tiedemann. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, page 7, 2016.

Chi-kiu Lo, Cyril Goutte, and Michel Simard. CNRC at SemEval-2016 Task 1: Experiments in Crosslingual Semantic Textual Similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 668–673, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1102.

Minh-Thang Luong and Christopher D Manning. Stanford Neural Machine Translation Systems for Spoken Language Domains. In *IWSLT (International Workshop on Spoken Language Translation)*, page 4, 2015.

Thang Luong, Hieu Pham, and Christopher D. Manning. Bilingual Word Representations with Monolingual Quality in Mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado, 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1521.

Bill MacCartney and Christopher D. Manning. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing - RTE '07*, page 193, Prague, Czech Republic, 2007. Association for Computational Linguistics. doi: 10.3115/1654536.1654575.

Bill MacCartney and Christopher D. Manning. An extended model of natural logic. In *Proceedings of the Eighth International Conference on Computational Semantics - IWCS-8 '09*, page 140, Tilburg, The Netherlands, 2009. Association for Computational Linguistics. ISBN 978-90-74029-34-6. doi: 10.3115/1693756.1693772.

Nitin Madnani and Bonnie J. Dorr. Generating Phrasal and Sentential Paraphrases: A Survey of Data-driven Methods. *Comput. Linguist.*, 36(3):341–387, September 2010. ISSN 0891-2017. doi: 10.1162/coli_a_00002.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online Dictionary Learning for Sparse Coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 689–696, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553463.

Andre Martins, Miguel Almeida, and Noah A. Smith. Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 617–622, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Austin Matthews, Waleed Ammar, Archna Bhatia, Weston Feely, Greg Hanneman, Eva Schlinger, Swabha Swayamdipta, Yulia Tsvetkov, Alon Lavie, and Chris Dyer. The CMU Machine Translation Systems at WMT 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 142–149, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3315.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. Cheap Translation for Cross-Lingual Named Entity Recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, September 2017. doi: 10.18653/v1/D17-1269.

Ryan McDonald, Slav Petrov, and Keith Hall. Multi-Source Transfer of Delexicalized Dependency Parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu, and Jungmee Lee. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, aug 2013.

Paul McNamee, Rion Snow, Patrick Schone, and James Mayfield. Learning Named Entity Hyponyms for Question Answering. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, June 1947. ISSN 1860-0980. doi: 10.1007/BF02295996.

Yashar Mehdad, Matteo Negri, and Marcello Federico. Towards Cross-Lingual Textual Entailment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 321–324, Los Angeles, California, USA, 2010.

George A. Miller. WordNet: A Lexical Database for English. *Commun. ACM*, 38 (11):39–41, November 1995. ISSN 0001-0782. doi: 10.1145/219717.219748.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, 3(4):235–244, December 1990. ISSN 0950-3846. doi: 10.1093/ijl/3.4.235.

Jeff Mitchell and Mirella Lapata. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429, November 2010. ISSN 03640213. doi: 10.1111/j.1551-6709.2010.01106.x.

Dragos Stefan Munteanu and Daniel Marcu. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4): 477–504, 2005.

Dragos Stefan Munteanu and Daniel Marcu. Extracting Parallel Sub-sentential Fragments from Non-parallel Corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 81–88, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220186.

Brian Murphy, Partha Talukdar, and Tom Mitchell. Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In *Proceedings of COLING 2012*, pages 1933–1950, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

M. Lynne Murphy. *Lexical Meaning*. Cambridge University Press, 2010.

Vinod Nair and Geoffrey E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA, 2010. Omnipress. ISBN 978-1-60558-907-7.

Ndapandula Nakashole and Raphael Flauger. Knowledge Distillation for Bilingual Dictionary Induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2497–2506, September 2017. doi: 10.18653/v1/D17-1264.

Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, December 2012. ISSN 00043702. doi: 10.1016/j.artint.2012.07.001.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. Semeval-2012 task 8: Cross-lingual textual entailment for content synchronization. In *Proceedings of 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*, pages 399–407. Association for Computational Linguistics, 2012.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. Semeval-2013 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 25–33, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

Peter Newmark. *Approaches to Translation*. Pergamon Press, Oxford; New York, 1981. ISBN 978-0-08-024603-1 978-0-08-024602-4 978-0-08-035603-7. OCLC: 6813775.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. Distinguishing Antonyms and Synonyms in a Pattern-based Neural Network. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 76–85, Valencia, Spain, 2017. Association for Computational Linguistics. doi: 10.18653/v1/E17-1008.

Eugene Albert Nida. *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. Brill Archive, 1964.

Sebastian Padó and Mirella Lapata. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199, June 2007. ISSN 0891-2017, 1530-9312. doi: 10.1162/coli.2007.33.2.161.

Sebastian Pado, Michel Galley, Dan Jurafsky, and Christopher D. Manning. Robust Machine Translation Evaluation with Entailment Features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 297–305, Suntec, Singapore, August 2009. Association for Computational Linguistics.

Despoina Panou. Equivalence in Translation Theories: A Critical Evaluation. *Theory and Practice in Language Studies*, 3(1):1–6, January 2013. ISSN 1799-2591. doi: 10.4304/tpls.3.1.1-6.

Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 113–120, July 2006. doi: 10.3115/1220175.1220190.

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

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. The Language Demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92, December 2014. ISSN 2307-387X. doi: 10.1162/tacl_a_00167.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. Adding Semantics to Data-Driven Paraphrasing. pages 1512–1522. Association for Computational Linguistics, 2015. doi: 10.3115/v1/P15-1146.

Yves Peirsman and Sebastian Padó. Semantic relations in bilingual lexicons. *ACM Transactions on Speech and Language Processing*, 8(2):1–21, November 2011. ISSN 15504875. doi: 10.1145/2050100.2050102.

Marco Pennacchiotti and Patrick Pantel. A Bootstrapping Algorithm for Automatically Harvesting Semantic Relations. In *Proceedings of the Fifth International Workshop on Inference in Computational Semantics (ICoS-5)*, 2006.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations.

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 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202.

MinhQuang Pham, Josep Crego, Jean Senellart, and François Yvon. 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, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1328.

Simone Paolo Ponzetto and Michael Strube. Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 192–199, New York City, USA, June 2006. Association for Computational Linguistics.

Reinhard Rapp. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics* -, page 320, Cambridge, Massachusetts, 1995. Association for Computational Linguistics. doi: 10.3115/981658.981709.

Mohammad Sadegh Rasooli and Michael Collins. Cross-Lingual Syntactic Transfer with Limited Resources. *Transactions of the Association for Computational Linguistics*, 5:279–293, December 2017. doi: 10.1162/tacl_a_00061.

Mohammad Sadegh Rasooli and Joel Tetreault. Yara Parser: A Fast and Accurate Dependency Parser. *arXiv:1503.06733 [cs]*, March 2015.

Philip Resnik. Mining the Web for Bilingual Text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 527–534, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. doi: 10.3115/1034678.1034757.

Philip Stuart Resnik. *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD Thesis, University of Pennsylvania, Philadelphia, PA, USA, 1993.

Jason Riesa and Daniel Marcu. Automatic Parallel Fragment Extraction from Noisy Data. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 538–542, Montréal, Canada, June 2012. Association for Computational Linguistics.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. Reasoning about Entailment with Neural Attention. *arXiv:1509.06664 [cs]*, September 2015.

Stephen Roller and Katrin Erk. Relations such as Hypernymy: Identifying and Exploiting Hearst Patterns in Distributional Vectors for Lexical Entailment. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2163–2172, Austin, Texas, November 2016. Association for Computational Linguistics.

Stephen Roller, Katrin Erk, and Gemma Boleda. Inclusive yet Selective: Supervised Distributional Hypernymy Detection. In *COLING*, pages 1025–1036, 2014.

Eleanor Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192–233, 1975. ISSN 1939-2222(Electronic),0096-3445(Print). doi: 10.1037/0096-3445.104.3.192.

Michael Roth and Shyam Upadhyay. Combining Discourse Markers and Cross-lingual Embeddings for Synonym–Antonym Classification. 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 3899–3905, June 2019. doi: 10.18653/v1/N19-1390.

A. Saboor and M. A. Khan. Lexical-semantic divergence in Urdu-to-English Example Based Machine Translation. In *2010 6th International Conference on Emerging Technologies (ICET)*, pages 316–320, October 2010. doi: 10.1109/ICET.2010.5638469.

Diana Santos. Lexical gaps and idioms in machine translation. In *Proceedings of the 13th Conference on Computational Linguistics* -, volume 2, pages 330–335, Helsinki, Finland, 1990. Association for Computational Linguistics. doi: 10.3115/997939.997996.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. Chasing Hypernyms in Vector Spaces with Entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers*, pages 38–42, Gothenburg, Sweden, 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-4008.

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. EVALution 1.0: An Evolving Semantic Dataset for Training and Evaluation of Distributional Semantic Models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 64–69, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4208.

Enrico Santus, Anna Gladkova, Stefan Evert, and Alessandro Lenci. The CogALex-V Shared Task on the Corpus-Based Identification of Semantic Relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 69–79. The COLING 2016 Organizing Committee, 2016.

M. Schuster and K.K. Paliwal. Bidirectional Recurrent Neural Networks. *Trans. Sig. Proc.*, 45(11):2673–2681, November 1997. ISSN 1053-587X. doi: 10.1109/78.650093.

Hinrich Schütze. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

Holger Schwenk and Matthijs Douze. Learning Joint Multilingual Sentence Representations with Neural Machine Translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2619.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *arXiv:1907.05791 [cs]*, July 2019.

Rico Sennrich and Biao Zhang. Revisiting Low-Resource Neural Machine Translation: A Case Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1021.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, August 2016. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. Adding Context to Semantic Data-Driven Paraphrasing. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 108–113, Berlin, Germany, August 2016a. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. Path-based vs. Distributional Information in Recognizing Lexical Semantic Relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 24–29. The COLING 2016 Organizing Committee, 2016b.

Vered Shwartz and Ido Dagan. CogALex-V Shared Task: LexNET - Integrated Path-based and Distributional Method for the Identification of Semantic Relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 80–85. The COLING 2016 Organizing Committee, 2016c.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. Improving Hypernymy Detection with an Integrated Path-based and Distributional Method. pages 2389–2398. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-1226.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. Hypernyms under Siege: Linguistically-motivated Artillery for Hypernymy Detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational*

Linguistics: Volume 1, Long Papers, pages 65–75, Valencia, Spain, April 2017. Association for Computational Linguistics.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411, Los Angeles, California, June 2010. Association for Computational Linguistics.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. Dirt Cheap Web-Scale Parallel Text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv:1702.03859 [cs]*, February 2017.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning Syntactic Patterns for Automatic Hypernym Discovery. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1297–1304. MIT Press, 2005.

Lucia Specia, Dhwaj Raj, and Marco Turchi. Machine Translation Evaluation Versus Quality Estimation. *Machine Translation*, 24(1):39–50, March 2010. ISSN 0922-6567. doi: 10.1007/s10590-010-9077-2.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André F. T. Martins. 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, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6451.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 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, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1150.

Jörg Tiedemann. Improved sentence alignment for movie subtitles. In *In Proceedings of RANLP, Borovets*, 2007.

Jörg Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Languages Resources Association (ELRA).

Marco Turchi and Matteo Negri. ALTN: Word Alignment Features for Cross-lingual Textual Entailment. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 128–132, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

P. D. Turney and S. M. Mohammad. Experiments with three approaches to recognizing lexical entailment. *Natural Language Engineering*, 21(03):437–476, May 2015. ISSN 1351-3249, 1469-8110. doi: 10.1017/S1351324913000387.

Peter Turney. A Uniform Approach to Analogies, Synonyms, Antonyms, and Associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 905–912. COLING 2008 Organizing Committee, 2008.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. Cross-lingual Models of Word Embeddings: An Empirical Comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1661–1670, Berlin, Germany, August 2016. Association for Computational Linguistics.

Shyam Upadhyay, Yogarshi Vyas, Marine Carpuat, and Dan Roth. Robust Cross-Lingual Hypernymy Detection Using Dependency Context. 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 607–618, June 2018. doi: 10.18653/v1/N18-1056.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

Jean-Paul Vinay and Jean-Louis Darbelnet. *Stylistique comparée du français et de l'anglais: méthode de traduction*. G.G. Harrap ; M. Didier, London; Toronto; Paris, 1958. OCLC: 713904127.

Ellen M. Voorhees. Query Expansion using Lexical-Semantic Relations. In Bruce W. Croft and C. J. van Rijsbergen, editors, *SIGIR '94*, pages 61–69. Springer London, 1994. ISBN 978-1-4471-2099-5.

Ivan Vulić. Cross-Lingual Syntactically Informed Distributed Word Representations. In *Proceedings of the 15th Conference of the European Chapter of the Association*

for Computational Linguistics, volume Volume 2, Short Papers, pages 408–414, Valencia, Spain, 2017. Association for Computational Linguistics. doi: 10.18653/v1/E17-2065.

Ivan Vulić and Marie-Francine Moens. Bilingual Distributed Word Representations from Document-Aligned Comparable Data. *Journal of Artificial Intelligence Research*, 55:953–994, April 2016. ISSN 1076-9757. doi: 10.1613/jair.4986.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. HyperLex: A Large-Scale Evaluation of Graded Lexical Entailment. *Computational Linguistics*, 43(4):781–835, September 2017. ISSN 0891-2017. doi: 10.1162/COLI_a_00301.

Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. Multilingual and Cross-Lingual Graded Lexical Entailment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4963–4974, Florence, Italy, July 2019. Association for Computational Linguistics.

Yogarshi Vyas and Marine Carpuat. Sparse Bilingual Word Representations for Cross-lingual Lexical Entailment. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1187–1197, June 2016. doi: 10.18653/v1/N16-1142.

Yogarshi Vyas and Marine Carpuat. Detecting Asymmetric Semantic Relations in Context: A Case-Study on Hypernymy Detection. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 33–43, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-1004.

Yogarshi Vyas and Marine Carpuat. Weakly Supervised Cross-lingual Semantic Relation Classification via Knowledge Distillation. In *EMNLP 2019*, Hong Kong, November 2019.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 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, June 2018. doi: 10.18653/v1/N18-1136.

Julie Weeds and David Weir. A General Framework for Distributional Similarity. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 81–88, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1119355.1119366.

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING 2014,*

the 25th International Conference on Computational Linguistics: Technical Papers, pages 2249–2259. Dublin City University and Association for Computational Linguistics, 2014.

Dekai Wu. Alignment. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition.*, pages 367–408. Chapman and Hall/CRC, 2010. ISBN 978-1-4200-8592-1.

Ruochen Xu and Yiming Yang. Cross-lingual Distillation for Text Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1130.

Yong Xu and Francois Yvon. Novel elicitation and annotation schemes for sentential and sub-sentential alignments of bitexts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*, pages 628–635, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).

Yadollah Yaghoobzadeh and Hinrich Schütze. Intrinsic Subspace Evaluation of Word Embedding Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 236–246, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1023.

Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*, pages 523–530, Toulouse, France, 2001. Association for Computational Linguistics. doi: 10.3115/1073012.1073079.

David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research - HLT '01*, pages 1–8, San Diego, 2001. Association for Computational Linguistics. doi: 10.3115/1072133.1072187.

Wen-tau Yih, Geoffrey Zweig, and John Platt. Polarity Inducing Latent Semantic Analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1212–1222, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

Savaş Yildirim and Tuğba Yildiz. Automatic Extraction of Turkish Hypernym-Hyponym Pairs From Large Corpus. In *Proceedings of COLING 2012: Demonstration Papers*, pages 493–500, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

Wenpeng Yin and Dan Roth. TwoWingOS: A Two-Wing Optimization Strategy for Evidential Claim Verification. In *Proceedings of the 2018 Conference on Empirical*

Methods in Natural Language Processing, pages 105–114, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1010.

Daniel Zeman and Philip Resnik. Cross-Language Parser Adaptation between Related Languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, page 8, 2008.

Yuming Zhai, Aurélien Max, and Anne Vilnat. 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, August 2018. Association for Computational Linguistics.

Yuming Zhai, Pooyan Safari, Gabriel Illouz, Alexandre Allauzen, and Anne Vilnat. Towards Recognizing Phrase Translation Processes: Experiments on English-French. *arXiv:1904.12213 [cs]*, April 2019.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J. Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. An Empirical Exploration of Curriculum Learning for Neural Machine Translation. *arXiv:1811.00739 [cs]*, November 2018.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. Curriculum Learning for Domain Adaptation in Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1903–1915, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1189.

Jiang Zhao, Man Lan, and Zheng-Yu Niu. ECNUCS: Recognizing Cross-lingual Textual Entailment Using Multiple Text Similarity and Text Difference Measures. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 118–123, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

A. Zouaq and R. Nkambou. Building Domain Ontologies from Text for Educational Purposes. *IEEE Transactions on Learning Technologies*, 1(1):49–62, January 2008. ISSN 1939-1382. doi: 10.1109/TLT.2008.12.