Weakly Supervised Cross-lingual Semantic Relation Classification via Knowledge Distillation

Yogarshi Vyas
Department of Computer Science
University of Maryland
yogarshi@cs.umd.edu

Marine Carpuat
Department of Computer Science
University of Maryland
marine@cs.umd.edu

Abstract

Words in different languages rarely cover the exact same semantic space. This work characterizes differences in meaning between words across languages using semantic relations that have been used to relate the meaning of English words. However, because of translation ambiguity, semantic relations are not always preserved by translation. We introduce a cross-lingual relation classifier trained only with English examples and a bilingual dictionary. Our classifier relies on a novel attention-based distillation approach to account for translation ambiguity when transferring knowledge from English to cross-lingual settings. On new English-Chinese and English-Hindi test sets, the resulting models largely outperform baselines that more naïvely rely on bilingual embeddings or dictionaries for cross-lingual transfer, and approach the performance of fully supervised systems on English tasks.

1 Introduction

Natural Language Processing (NLP) often uses translation lexicons for projecting models and data from one language to another under the assumption that words and their translations in these lexicons are synonyms (Mayhew et al., 2017; Tsvetkov et al., 2014). However, translation lexicons include semantic relations other than synonymy in practice, as can be seen (Table 1) in examples drawn from the MUSE dictionary (Lample et al., 2018). Peirsman and Padó (2011) show that distributional translation lexicons contain hyponyms and co-hyponyms, and that treating all translations as synonyms hurts cross-lingual projection performance. In the Paraphrase Database (Ganitkevitch et al., 2013), Pavlick et al. (2015) find that the diversity of semantic relations discovered in word-aligned parallel corpora yields paraphrases that span the lexical relations defined in the natural logic framework of MacCartney and Manning (2009). These non-synonymous translations are not just noise. They can be found even in high-quality parallel corpora, since the strategies used by professional translators to deal with words that do not have a direct equivalent in the target language include replacement by near-synonyms, hypernyms or negated antonyms (Baker, 2011; Venuti, 2012; Chesterman, 2016).

Lexicon Entry	Semantic Relation
writer, रचनाकार	*writer* is more specific than रचनाकार (creator)
council, मंत्रिपरिषड	*council* is more general than मंत्रिपरिषड (council of ministers)
father, चचा	*father* is mutually exclusive to चचा (father's brother)

Table 1: Semantic relations between word pairs in an English-Hindi lexicon (Lample et al., 2018)

In this work, we classify semantic relations between words in different languages. Given a word pair (water, पय), the classification task is to select one of the five entailment classes (Figure 1) defined under the natural logic framework of MacCartney and Manning (2009). This cross-lingual task cannot be solved by translation, as translation does not preserve semantic relations. We also cannot assume that labeled examples exist for all language pairs and learning from English labeled examples is complicated by translation ambiguity (Figure 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, 2016a,b) (Section 3). 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 dic-

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5285–5296, Hong Kong, China, November 3–7, 2019.

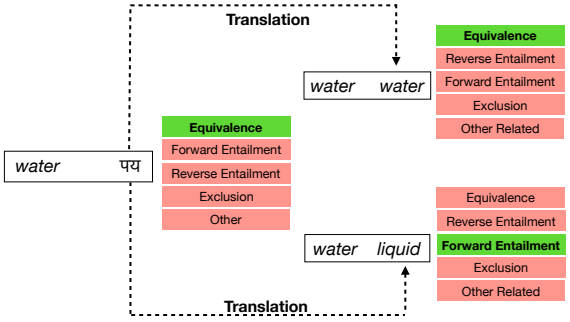

Figure 1: On the left, we illustrate cross-lingual semantic relation classification: given the pair (*water*, पय) as input, the task is to select the `Equivalence` class (in bold/green) from the five possible relations. On the right, we show that semantic relations change by translation. पय translates to *liquid* and *water*, and their respective semantic relations with *water* differ.

tionaries (Hinton et al., 2014) (Section 4). We collect and release MULTILEXREL, a crowd-sourced benchmark to evaluate models for this task on a high-resource (English-Chinese) and a low-resource (English-Hindi) language pair (Section 5). Experiments show that BILEXNET substantially outperforms translation baselines and approaches the performance of a fully supervised English semantic relation classifier (Section 7). Code for BILEXNET and the MULTILEXREL dataset are available at `https://github.com/yogarshi/BiLexNet/`.

2 Related Work

Cruse (1986) describes how lexical relations can be organized into congruence relations of identity (synonymy: dog-canine), inclusion (hyponymy: dog-animal), overlap (compatibility: dog-pet), disjunction (incompatibility: cat-dog), and antonyms (open-shut), and how these relations can be used to characterize differences in meaning. Variations on these fundamental relations have been used within semantic networks such as WordNet (Fellbaum, 1998), or as the basis of a framework for inference without formal logic representations (MacCartney and Manning, 2009). Recent work on semantic relation prediction largely focuses on a single relation between words in the same language (mostly English) (Nastase et al., 2013; Vulić and Mrkšić, 2018; Glavaš and Ponzetto, 2017; Ono et al., 2015). 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).

In cross-lingual settings, studies of semantic relations between words are mostly limited to the translation equivalence relation. Dictionary induction aims to automatically discover words that are translations of each other using monolingual or comparable corpora (Rapp, 1995; Fung and Yee, 1998; Irvine and Callison-Burch, 2017). The task is typically framed as unsupervised learning, and models rely on distributional properties to discover words that have the same meaning in the two languages. Recently, dictionary induction has also been used to evaluate multilingual word embeddings (Lample et al., 2018; Artetxe et al., 2018; Søgaard et al., 2018), leading to significant advances on the state-of-the-art.

However, Peirsman and Padó (2011) show that automatically induced bilingual lexicons exhibit multiple semantic relations including not only synonymy, but also hypernymy and co-hyponymy. This has prompted work on identifying hypernymy in cross-lingual settings (Vyas and Carpuat, 2016; Upadhyay et al., 2018). This paper broadens the scope of relations studied in cross-lingual settings, and addresses for the first time the task of distinguishing between multiple semantic relations for words in different languages.

Previous work studying semantic relations in multiple languages has focused on the different task of cross-lingual transfer. In such settings, the focus is on identifying semantic relations between two words in the *same* language, without training data in that language. There are broadly two strategies to solve this problem. One line of work (Glavaš and Vulić, 2018; Mrkšić et al., 2017) uses model transfer, where a single model is trained on data from a high-resource language, and is then ported to the target language using cross-lingual embeddings. In contrast, Roth and Upadhyay (2019) translate training data from English into a target language using a combination of unsupervised cross-lingual embeddings (Artetxe et al., 2018) and monolingual information from the target language.

Our approach for cross-lingual semantic relation classification builds on the monolingual classifier LEXNET (Shwartz and Dagan, 2016a,b), which 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 proposing to model cross-lingual relations using lexico-syntactic paths from both languages.

Finally, our training procedure uses knowledge distillation (Hinton et al., 2014) to alleviate the lack of annotated cross-lingual pairs. 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.

3 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/ is a type of Y), `Backward Entail` (X is more general than/encompasses Y), `Exclusion` (X is mutually exclusive with/is 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, 2016a,b) 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 context (Hearst, 1992; Snow et al., 2004; Shwartz et al., 2016). For classification, $\mathbf{v}_{xy}$ is input to a multi-class classifier, parameterized as a feed-forward neural network with a single hidden layer.

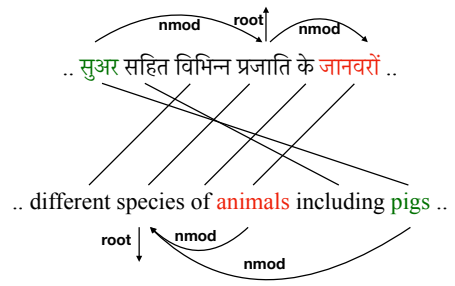

Figure 2: The English path between *animals* and *pigs* has three edges: [`X/NOUN/nmod/>`, `species/NOUN/root/^`, and `Y/NOUN/nmod/>`]. The path between animals and सुअर is defined as a combination of the English path and the Hindi path between जानवरों and सुअर.

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

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

3.1 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 2 for an example of the path connecting *pigs* and *animals*). The vector representation of each edge is formed by the concatenation of 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 word-aligned parallel corpus (Figure 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.

4 Weakly Supervised Training via Knowledge Distillation

Cross-lingual examples that would enable 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 $\hat{\mathbf{l}}_i$:

$$L_1 = -\sum_{j=1}^{c} l_{ij} \log \hat{l}_j^{e\to e} \tag{2}$$

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 obtained 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 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 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.

$$score((x_{e;i}, t_{ik}), \mathbf{l}_i) = f([\mathbf{x}_i; \mathbf{t}_{ik}; \mathbf{l}]) \tag{3}$$

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

$$\mathbf{v}'_{x_{e;i}y_{e;i}} = \sum_{k=1}^{n} p(t_{ik}) \mathbf{v}_{x_{e;i}t_{ik}} \tag{4}$$

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

$$L_2 = -\sum_{j=1}^{c} [(1-\alpha)\, l_{ij} \log \hat{l}_j^{e\to f} + \alpha\, \hat{l}_j^{e\to e} \log(\frac{\hat{l}_j^{e\to f}}{\hat{l}_j^{e\to e}})] \tag{5}$$

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

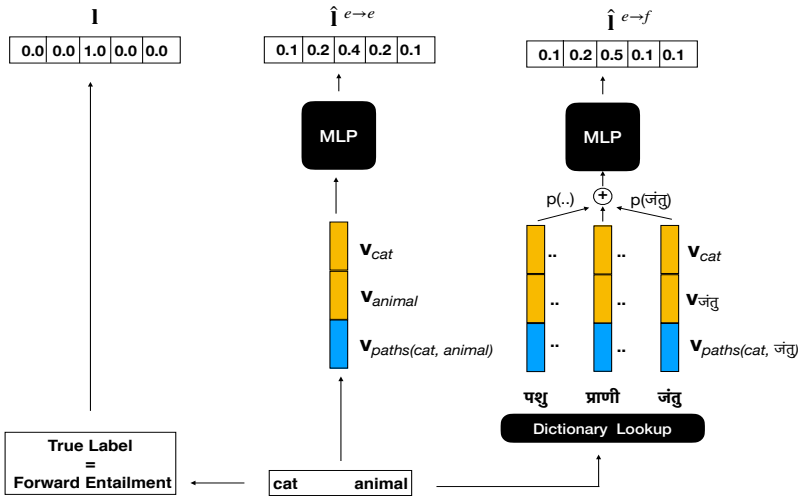

Figure 3: Illustration of weakly supervised training: For a given English example (*cat*, *animal*), we generate predictions $\hat{\mathbf{l}}^{e \to e}$ using the monolingual English teacher model. Simultaneously, we also generate predictions $\hat{\mathbf{l}}^{e \to f}$ using the cross-lingual *student* 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($\hat{\mathbf{l}}^{e \to e}$, l) + CROSS-ENTROPY($\hat{\mathbf{l}}^{e \to f}$, l) + KL-DIVERGENCE($\hat{\mathbf{l}}^{e \to f}$, $\hat{\mathbf{l}}^{e \to e}$).

that aims to measure how well the cross-lingual model M_{ef} predicts 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 monolingual 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 with a temperature parameter τ.

5 MULTILEXREL: A Dataset for Cross-lingual Semantic Relations

Existing resources containing annotated cross-lingual lexical relations are limited in scope, quality, and quantity. For instance, in previous work (Upadhyay et al., 2018), we provide datasets annotated cross-lingual hypernymy, but do not consider other relations. On the other hand, 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 are likely to be noisy as these resources are created in a semi-automatic way.

In this work, we crowdsource[1] 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 3. We leverage monolingual annotations

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

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 (Hovy et al., 2013), a Bayesian model that estimates the trustworthiness of annotators and accordingly assigns a label to each instance. The distributions of the five relations for each language pair are shown in Table 2. 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.[2]

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

[2] A detailed description is in the appendix.

6 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.

6.1 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 `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 (Levy et al., 2015), we ensure a lexical split where the English words in the test data are not seen in the training data. This makes the task challenging as it prevents the 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.[3] We use FastText bilingual embeddings (Bojanowski et al., 2017).[4] We extract English paths for the monolingual model from a dump of the English Wikipedia.[5] Cross-lingual paths are extracted from a random sample of the WMT18 parallel corpora[6] 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,[7] and the Jieba word segmenter for Chinese.[8]

6.2 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, 2015) 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.[9]

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

[3] https://www.mdbg.net/chinese/dictionary?page=cc-cedict

[4] https://fasttext.cc/docs/en/aligned-vectors.html

[5] https://dumps.wikimedia.org/enwiki/

[6] http://statmt.org/wmt18/translation-task.html

[7] https://github.com/anoopkunchukuttan/indic_nlp_library

[8] https://github.com/fxsjy/jieba

[9] We re-implement the LEXNET model and verify its accuracy by replicating results on the CogALex dataset.

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 2.[10]

BILEXNET (NO DISTILLATION) A simple strategy for cross-lingual transfer consists of 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 (Klementiev et al., 2012; Guo et al., 2015, *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 aims to answer 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 dataset. 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 STM in a new setting to predict the semantic relation between two words in different languages.

We use the official implementation of STM[11] with the same bilingual embeddings used by BILEXNET. We tune three hyperparameters on the validation set (a) The size of the specialized tensors {100, 200, 300, 500} (b) The number of specialization functions {3, 5, 7}, and (c) The learning rate {0.0001, 0.0003}. Default values are used for all other hyper-parameters.

7 Results

Tables 3 summarizes results obtained 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-only 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 BASELINE, 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 6) 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 ($\sim$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 $\sim$9 points worse than the full BILEXNET model, losing on both

[10] We also experimented with a voting based approach for combination, but it generally performed worse.

[11] 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 (Glavaš and Vulić, 2018)	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 3: Precision (P), Recall (R) and F1-score (F) for BILEXNET and contrastive baselines on the two MULTILEXREL test sets. All configurations (except STM) 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 attention.

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., 2016). 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 cross-lingual 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.

8 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.

Performance Per Class We break down the performance of the BILEXNET model per target relation (Table 4). 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, 2016a): 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 Co-

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 4: Per-class F1 scores for median En-Hi and En-Zh BILEXNET model and the ENLEXNET model.

gALex shared task (Santus et al., 2016), where the first part of the task is to eliminate *completely unrelated* pairs, before predicting relations on the remaining pairs. 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 3 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.

9 Conclusion

This work contributes data and models to the task of classifying semantic relations between words in different languages with only monolingual English supervision. 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 proposed 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.

Acknowledgements

The authors would like to thank members of the CLIP Lab and the anonymous reviewers for feedback on previous versions of this paper. This material is based upon work supported by the National Science Foundation under Award No. 1750695.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Mona Baker. 2011. *In other words: A coursebook on translation*. Routledge.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Francis Bond and Ryan Foster. 2013. 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. Association for Computational Linguistics.

Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. 2018a. 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.

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

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

Andrew Chesterman. 2016. *Memes of translation: The spread of ideas in translation theory*, volume 123. John Benjamins Publishing Company.

David Alan Cruse. 1986. *Lexical semantics*. Cambridge university press.

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

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Pascale Fung and Lo Yuen Yee. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 414–420, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Goran Glavaš and Simone Paolo Ponzetto. 2017. 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. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 2018. 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. Association for Computational Linguistics.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. 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.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2014. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. 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.

Ann Irvine and Chris Callison-Burch. 2017. A Comprehensive Analysis of Bilingual Lexicon Induction. *Computational Linguistics*, 43(2):273–310.

Diederick P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*.

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

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. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster*

Sessions, pages 177–180. Association for Computational Linguistics.

S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

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

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

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. 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. Association for Computational Linguistics.

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

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eighth International Conference on Computational Semantics - IWCS-8 '09*, page 140, Tilburg, The Netherlands. Association for Computational Linguistics.

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

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. 2013. 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.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association of Computational Linguistics*, 5:309–324.

Ndapandula Nakashole and Raphael Flauger. 2017. Knowledge Distillation for Bilingual Dictionary Induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2497–2506.

Vivi Nastase, Preslav Nakov, Diarmuid Ó Séaghdha, and Stan Szpakowicz. 2013. Semantic Relations Between Nominals. *Synthesis Lectures on Human Language Technologies*, 6(1):1–119.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word Embedding-based Antonym Detection using Thesauri and Distributional Information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 984–989. Association for Computational Linguistics.

Patrick Pantel and Marco Pennacchiotti. 2006. 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.

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

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

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

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

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara Parser: A Fast and Accurate Dependency Parser. *CoRR*, abs/1503.06733.

Michael Roth and Shyam Upadhyay. 2019. 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.

Enrico Santus, Anna Gladkova, Stefan Evert, and Alessandro Lenci. 2016. 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.

Vered Shwartz and Ido Dagan. 2016a. 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.

Vered Shwartz and Ido Dagan. 2016b. 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.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2389–2398. Association for Computational Linguistics.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2004. 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.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788. Association for Computational Linguistics.

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

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Peter Turney. 2008. 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.

Shyam Upadhyay, Yogarshi Vyas, Marine Carpuat, and Dan Roth. 2018. 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.

Lawrence Venuti. 2012. *The translation studies reader*. Routledge.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising Word Vectors for Lexical Entailment. 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 1134–1145, New Orleans, Louisiana. Association for Computational Linguistics.

Yogarshi Vyas and Marine Carpuat. 2016. 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.

Ruochen Xu and Yiming Yang. 2017. 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.

Wen-tau Yih, Geoffrey Zweig, and John Platt. 2012. 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. Association for Computational Linguistics.