Cross-lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need!

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
http://www.ukp.tu-darmstadt.de

Abstract

Argumentation mining (AM) requires the identification of complex discourse structures and has lately been applied with success monolingually. In this work, we show that the existing resources are, however, not adequate for assessing cross-lingual AM, due to their heterogeneity or lack of complexity. We therefore create suitable parallel corpora by (human and machine) translating a popular AM dataset consisting of persuasive student essays into German, French, Spanish, and Chinese. We then compare (i) annotation projection and (ii) bilingual word embeddings based direct transfer strategies for cross-lingual AM, finding that the former performs considerably better and almost eliminates the loss from cross-lingual transfer. Moreover, we find that annotation projection works equally well when using either costly human or cheap machine translations. Our code and data are available at http://github.com/UKPLab/coling2018-xling_argument_mining.

1 Introduction

Argumentation mining (AM) is a fast-growing research field with applications in discourse analysis, summarization, debate modeling, and law, among others (Peldszus and Stede, 2013a). Recent studies have successfully applied computational methods to analyze monological argumentation (Wachsmuth et al., 2017; Eger et al., 2017). Most of these studies view arguments as consisting of (at least) claims and premises—and so do we in this work. Thereby, our focus is on token-level *argument component extraction*, that is, the segmentation and typing of argument components.

AM has thus far almost exclusively been performed *monolingually*, e.g. in English (Mochales-Palau and Moens, 2009), German (Eckle-Kohler et al., 2015), or Chinese (Li et al., 2017). Working only monolingually is problematic, however, because AM is a difficult task even for humans due to its dependence on background knowledge and parsing of complex pragmatic relations (Moens, 2017). As a result, acquiring (high-quality) datasets for new languages comes at a high cost—be it in terms of training and/or hiring expert annotators or querying large crowds in crowd-sourcing experiments. It is thus of utmost importance to train NLP systems in AM that are capable of going cross-language, so that annotation efforts do not have to be multiplied by the number of languages of interest. This is in line with current trends in NLP, which increasingly recognize the possibility and the necessity to work cross-lingually, be it in part-of-speech tagging (Zhang et al., 2016), dependency parsing (Agic et al., 2016), sentiment mining (Chen et al., 2016; Zhou et al., 2016), or other fields.

In this work, we address the problem of cross-lingual (token-level) AM for the first time. We initially experiment with available resources in English, German, and Chinese. We show that the existing datasets for analyzing argumentation are not suitable for assessing cross-lingual component extraction due to their heterogeneity or lack of complexity. Given this scarcity of homogeneously annotated high-quality large-scale datasets across different languages, our first contribution is to (1) provide a fully parallel (en-de), *human-translated* version of one of the most popular current AM datasets, namely, the English dataset of persuasive student essays published by Stab and Gurevych (2017).

Proceedings of the 27th International Conference on Computational Linguistics, pages 831–844
Santa Fe, New Mexico, USA, August 20-26, 2018.

We then (2) *machine translate* the 402 student essays into German, Spanish, French, and Chinese. Both our human and machine translations contain argumentation annotations, in the form of either human annotations or automatically projected annotations. Our experiments indicate that both the translations and the projected annotations are of very high quality, cf. examples in Table 2.

Besides contributing new datasets, (3) we perform the first evaluations of cross-lingual (token-level) AM, based on suitable adaptations of two popular cross-lingual techniques, namely, *direct transfer* (McDonald et al., 2011) and *projection* (Yarowsky et al., 2001). We find that projection works considerably better than direct transfer and almost closes the cross-lingual gap, i.e., cross-lingual performance is almost on par with in-language performance when we use parallel data and project annotations to the target language. This holds both for human (translated, HT) parallel data, which is costly to obtain, and machine translated (MT) parallel data, which is very cheap to obtain for dozens of high-resource languages.

Our findings imply that current neural MT has reached a level where it can act as a substitute for costly (non-expert) HT even for problems that operate on the fine-grained token-level.

2 Related Work

In what follows, we briefly summarize the works that most closely relate to our current research.

Argumentation Mining AM seeks to automatically identify argument structures in text and has received a lot of attention in NLP lately. Existing approaches focus, for instance, on specific subtasks like argument unit segmentation (Ajjour et al., 2017), identifying different types of argument components (Mochales-Palau and Moens, 2009; Habernal and Gurevych, 2017), recognizing argumentative discourse relations (Nguyen and Litman, 2016) or extracting argument components and relations end-to-end (Eger et al., 2017). However, most of these approaches are specifically designed for English and there are only few resources for other languages. For German, a few datasets annotated according to the claim-premise scheme are available (Eckle-Kohler et al., 2015; Liebeck et al., 2016). Furthermore, Peldszus and Stede (2015) annotated a small German dataset with claims and premises and translated it to English subsequently. There are very few works studying AM in other languages, e.g. Basile et al. (2016) for Italian, Li et al. (2017) for Chinese and Sardianos et al. (2015) for Greek. There are also two recent papers addressing some form of cross-linguality: Aker and Zhang (2017) map argumentative sentences from English to Mandarin using machine translation in comparable Wikipedia articles. Sliwa et al. (2018) create corpora in Balkan languages and Arabic by labeling the English side of corresponding parallel corpora on the sentence level and then using the same label for the target sentences. In contrast to these works, we work on the more challenging token-level. Moreover, we actually train classifiers for language transfer rather than only creating annotated data in other languages based on parallel data.

As mentioned, we focus on *cross-lingual component extraction*, that is, the segmentation and typing of argument components in a target language (L2), while having only annotated source language (L1) data. We operate on token-level by labeling each token with a BIO label plus its respective component type. The BIO label marks the start, continuation and end of specific argument components. Examples are given in Tables 2 and 3.

Cross-lingual sequence tagging POS tagging and named-entity recognition (NER) are standard tasks in NLP. In recent years, there is increased interest not only in evaluating POS and NER models within multiple individual languages (Plank et al., 2016), but also cross-lingually (Zhang et al., 2016; Tsai et al., 2016; Mayhew et al., 2017; Yang et al., 2017). Two standard approaches are *projection* (Yarowsky et al., 2001; Das and Petrov, 2011; Täckström et al., 2013) and *direct transfer* (Täckström et al., 2012; Zhang et al., 2016). Projection uses parallel data to project annotations from one language to another. In contrast, in direct transfer, a system is trained in L1 using language independent or shared features and applied without modification to L2.

While these approaches are typically *unsupervised*, i.e., they assume no annotations in L2, there are also *supervised* cross-lingual approaches based on multi-task learning (Cotterell and Heigold, 2017; Yang et al., 2017; Kim et al., 2017). These assume small training sets in L2, and a system trained on them is regularized by a larger amount of training data in L1. In our work, we only consider unsupervised

Name	Docum.	Tokens	Sentences	Major Cl.	Cl.	Prem.	Genre	Lang.
MTX	112	8,865 (en)	449	-	112	464	short texts	en, de
CRC	315	21,858	957	135	1,415	684	reviews	zh [en]
PE	402	148,186 (en)	7,141	751	1,506	3,832	persuasive essays	en [de, fr, es, zh]

Table 1: Statistics for datasets used in this work. Languages in brackets added by the current work.

Orig-EN	In the end , I think [**any great success need great work not great luck**] , even though [luck is one factor in reaching goal] but [*its impact is extraneous and we must not reckon on luck in our plans*] .
HT-DE-HumanAnno	Schließlich denke ich , dass [**jeder große Erfolg auf harter Arbeit statt Glück beruht**] , obwohl [Glück ein Faktor in der Erreichung des Ziels ist] , jedoch [*ist dessen Auswirkung unwesentlich und wir sollten uns nicht in unseren Projekten auf unser Glück verlassen*] .
HT-DE-ProjAnno	Schließlich denke ich , dass [**jeder große Erfolg auf harter Arbeit statt Glück beruht , obwohl Glück**] [ein Faktor in der Erreichung des Ziels ist] , jedoch [*ist dessen Auswirkung unwesentlich und wir sollten uns nicht in unseren Projekten auf unser Glück verlassen*] .
MT-DE-ProjAnno	Am Ende denke ich , dass [**jeder große Erfolg große Arbeit erfordert , nicht viel Glück**] , auch wenn [Glück ein Faktor beim Erreichen des Ziels] [*ist , aber seine Auswirkungen sind irrelevant und wir dürfen nicht mit Glück in unseren Plänen rechnen*] .
MT-ES-ProjAnno	Al final , creo que [**cualquier gran éxito requiere un gran trabajo y no mucha suerte**] , aunque la [suerte es un factor para alcanzar el objetivo] , pero [*su impacto es extraño y no debemos tener en cuenta la suerte en nuestros planes*] .
MT-FR-ProjAnno	En fin de compte , je pense que [**tout grand succès a besoin d' un bon travail , pas de chance**] , même si la [chance est un facteur d' atteinte de l' objectif] , mais [*son impact est étranger et nous ne devons pas compter sur la chance dans nos plans*] .
MT-ZH-ProjAnno	最后, 我认为[**任何伟大的成功都需要伟大的工作, 而不是运气好**] , 即使 [运气是达成目标的一个因素] , [但其影响是无关紧要的, 我们不能算计划中的运气] 。

Table 2: Human-annotated English sentence in the PE dataset as well as translations with human-created and projected annotations. Major claims in bold, claims underlined, premises in italics. HT/MT =human/machine translation.

language adaptation, because it is the most realistic cross-lingual scenario for AM, as it may be costly to even produce small amounts of training data in many different languages.

Most cross-lingual sequence tagging approaches address POS tagging, and only few are devoted to NER (Mayhew et al., 2017; Tsai et al., 2016), aspect-based sentiment classification (Lambert, 2015), or even more challenging problems such as discourse parsing (Braud et al., 2017). While POS tagging and NER are in some sense very similar to AM, namely, insofar as both can be modeled as sequential tagging of tokens, there are also important differences. For example, in POS tagging and NER, the label for a current token usually strongly depends on the token itself plus some local context. This strong association between label, token and local context is largely absent in AM, causing some models that perform well on POS and NER to fail blatantly in AM.[1]

Cross-lingual Word Embeddings are the (modern) basis of the direct transfer method. As with monolingual embeddings, there exists a veritable zoo of different approaches, but they often perform very similarly in applications (Upadhyay et al., 2016) and seemingly very different approaches are oftentimes also equivalent on a theoretical level (Ruder et al., 2017).

3 Data

We chose three freely available datasets: a small parallel German-English dataset, and considerably larger English and Chinese datasets using (almost) the same inventory of argument types, which we therefore assumed to be adequate for cross-lingual experiments. We translated the two last named monolingual datasets in other languages, described below. Statistics for all datasets are given in Table 1.

[1]E.g., we had tried out a word embedding based HMM model (Zhang et al., 2016) in initial experiments but found it to perform below our random baseline. The apparent reason is that an HMM cannot deal with long-range dependencies that abound in AM.

Orig-ZH	几次入住中国大饭店, [**感觉都非常不错**] , [新开的豪华阁酒廊非常棒] , [饮料丰富] , [食物也很好吃] , [服务也非常的棒] , [尤其特别感谢杨雪峰和张东静, 他们非常贴心]
MT-EN-ProjAnno	Several times staying at China World Hotel , [**I feel very good**] , the [newly opened Horizon Club Lounge is great] , [rich drinks] , [food is also very good] , [very good service] , [*especially thanks to Yang Xuefeng and Zhang Dongjing , they are very caring*]

Table 3: Review from CRC corpus as well as English machine translation with projected annotations. Major claims in bold, claims underlined, premises in italics (ZH: regular font).

3.1 Microtexts (MTX)

Peldszus and Stede (2015) annotated 112 German short texts (six or less sentences) written in response to questions typically phrased like "Should one do X". These were annotated according to a version of Freeman's theory of argumentation macro-structure (Peldszus and Stede, 2013b). Each microtext consists of one (central) claim and several premises. As opposed to our other datasets, MTX has no "O" (non-argumentative) tokens and no major claims. The German sentences have been professionally translated to English, making this the first parallel corpus for AM in English and German.

3.2 Chinese Review Corpus (CRC)

Li et al. (2017) created the only large-scale argument mining dataset in Chinese, freely available and with annotations on component level according to the claim-premise scheme (Stab and Gurevych, 2017). We thus chose to include this dataset in our experiments, despite differences in the domain of the annotated texts. Li et al. (2017) used crowdsourcing to annotate Chinese hotel reviews from *tripadvisor.com* with four component types (major claim, claim, premise, premise supporting an implicit claim). We consider only those components with direct overlap with the components used by Stab and Gurevych (2017), thus considering components labeled as "premise supporting an implicit claim" as non-argumentative. We applied the CRF-based Chinese word segmenter by Tseng et al. (2005) to split Chinese character streams into tokens. Furthermore, we only use the "Easy Reviews Corpus" from Li et al. (2017). The remaining part of the corpus are isolated sentences from reviews with low overall inter-annotator agreement, which we ignored. An example from CRC can be found in Table 3.

3.3 A Large-Scale Parallel Dataset of Persuasive Essays (PE)

Stab and Gurevych (2017) created a dataset of persuasive essays written by students on *essaysforum.com*. These are about controversial topics such as "competition or cooperation—which is better?". To obtain a human-translated parallel version of this dataset, we asked seven native speakers of German with an attested strong competence in English (all students or university employees) to translate the 402 student essays in the PE corpus sentence-by-sentence. As only requirement, we asked the translators to retain the argumentative structure in their translations: i.e., the translation of an argument component should be connected and not contain non-argumentative tokens. Since German has a freer word order compared to English, this requirement can in virtually all cases be easily fulfilled without producing awkward sounding German translations. Each essay was translated by exactly one translator. Besides translating the essays, we also asked the translators to annotate argument boundaries so that the original mark-up is preserved in the translations. The translators took about 40min on average to translate one essay and indicate the argument structures. Thus, they required about 270 hours to translate the whole PE corpus into German, and the resulting overall cost was roughly 3,000 USD. The motivations to ask translators to translate argument components contiguously were that (i) all monolingual AM datasets we know of have contiguous components, (ii) transfer would have been naturally hampered had components in the source language been contiguous but not in the target language, at least for methods such as direct transfer.[2]

[2]We note that even professional translations typically differ from original, non-translated texts because they retain traces of the source language (Rabinovich et al., 2017). We thus speculate that our reported results are probably slightly upward biased compared to a situation where the test data consists of original German student essays. This latter situation would have been much more costly to produce, in any way: it would have required retrieval (and, if necessary, creation) of original student essays in German as well as induction of all subsequent annotation mark-up.

To obtain further parallel versions of the PE data, we also *automatically* translated them into German, French, Spanish, and Chinese using Google Translate. Of course, we cannot make any demands on how Google Translate translates text into other languages but noticed that it has a tendency to stay rather close to the original text, but nevertheless has a very high perceived quality of translation. We automatically projected argument structures from the English text to the machine translations using our projection algorithm described in §4. It took few hours to automatically translate the PE corpus into the four languages. Examples of the data as well as the human and machine translations can be found in Table 2. Even though we also provide translations of PE in French, Spanish and Chinese, our primary focus in our experiments below is on the languages for which we have gold (human-created) data, i.e., EN↔DE (for PE and MTX) as well as EN↔ZH.

4 Approaches

In what follows, we describe our adaptations of direct transfer and projection to the AM task. Direct transfer focuses on the source language and trains on human-created L1 data as well as human-created L1 labels. In contrast, during training, projection operates directly on the language of interest, viz., L2. This comes at a cost: the labels in L2 are noisy, because they are projected from L1, which is an error-prone process. The success of projection can therefore be expected to largely depend on the quality of this transfer step. Projection makes stronger assumptions than direct transfer: it requires parallel data.[3] When the parallel data is induced via machine translation, then a second source of noise for projection is the 'unreliable' L2 input training data.

Direct Transfer Here, we directly train a system on bilingual representations, which in our case come in the form of bilingual word embeddings. To retain some freedom over the choice and parameters of our word embeddings, we choose to train them ourselves instead of using pre-trained ones. For EN↔DE we induce bilingual word embeddings by training BIVCD (Vulic and Moens, 2015) and BISKIP models (Luong et al., 2015) on >2 million aligned sentences from the Europarl corpus (Koehn, 2005). BIVCD concatenates bilingually aligned sentences (or documents), randomly shuffles each concatenation and trains a standard monolingual word embedding technique on the result; here, we use the word2vec skip-gram model (Mikolov et al., 2013). BIVCD was shown to be competitive to more challenging approaches in Upadhyay et al. (2016). BISKIP is a variant of the standard skip-gram model which predicts mono- and cross-lingual contexts. It requires word alignments between parallel sentences and we use fast-align for this (Dyer et al., 2013). For EN↔ZH we train the same models on the UN corpus (Ziemski et al., 2016), which comprises >11 million parallel sentences. We train embeddings of sizes 100 and 200.

Projection To implement projection for the problem of token-level AM, we proceed as follows. We take our human-labeled L1 data and align it with its corresponding parallel L2 data using fast-align. Once we have word level alignment information, we consider for each argument component $c(s)$ in L1 of type a (e.g., MajorClaim, Claim, Premise) with consecutive words $s_1, \ldots, s_N$: the word t_1 with smallest index in the corresponding L2 sentence that is aligned to some word in $s_1, \ldots, s_N$, and the analogous word t_{-1} with largest such index. We then label all the words in the L2 sentence between t_1 and t_{-1} with type a, using a correct BIO structure, resulting in $c(t)$. We repeat this process for all the components within a sentence in L1 and for all sentences. In case of collision, e.g., if two components in L2 would overlap according to the above-described strategies, we simply increment the beginning counter of one of the components until they are disjoint. If our above strategy fails, i.e., $c(s)$ cannot be projected, e.g., because the words in an L1 component are not aligned to any words in L2, then we simply ignore the projection of $c(s)$ to the L2 sentence, labeling the corresponding words in L2 as non-argumentative instead. We think of this projection strategy as naive because we do not do much to resolve conflicts and instead trust the quality of the alignments and that the subsequent systems trained on the projected data are capable of gracefully recovering from noise in the projections.

[3]Thus, direct transfer is potentially the cheaper approach, even though it also requires bilingual word embeddings, which themselves are based on some form of bilingual signal, e.g., parallel sentence- or word-level data.

5 Experiments

We perform token-level sequence tagging. Our label space is $\mathcal{Y} = \{\text{B,I}\} \times T \cup \{\text{O}\}$ where T is the set of argument types, comprising "claim", "premise", and (if applicable) "major claim".

5.1 Experimental Setup

To perform token-level sequence tagging, we implement a standard bidirectional LSTM with a CRF layer as output layer in TensorFlow. The CRF layer accounts for dependencies between successive labels. We represent words by their respective embeddings. In addition to this word-based information, we also allow the model to learn a character-based representation (via another LSTM) and concatenate this learned representation to the word embedding. Our model is essentially the same as the ones proposed by Ma and Hovy (2016) and Lample et al. (2016); it is also a state-of-the-art model for monolingual AM (Eger et al., 2017). We name it BLCRF+char, when character information is included, and BLCRF when disabled. For all experiments, we use the same architectural setup: we use two LSTM hidden layers with 100 hidden units each. We train for 50 epochs using a patience of 10. We apply dropout on the embeddings as well as on the LSTM units. On character-level, we also use a bidirectional LSTM with 50 hidden units and learn a representation of size 30. As evaluation measure we choose macro-F1 as implemented in scikit-learn (Pedregosa et al., 2011).

Baseline A simple baseline to test successful learning is to choose the majority label in the test data. However, this performs particularly poorly on token-level and for our chosen evaluation metric. We therefore choose a more sophisticated baseline. We first split our datasets by sentences and then compute a probability distribution of how likely each argument component appears in a sentence. At test time, we again split the test data by sentences and then label each token in the test sentence with a randomly drawn argument component (according to the calculated probability distribution on train/dev sets). We label all the tokens in the sentence with the drawn argument component type, keeping valid BIO structure. We label the last token (which is typically a punctuation symbol) with the "O" label in PE and CRC. In essence, our baseline is a random baseline, but has some basic prior knowledge of the BIO format.

Train/dev/test splits For the PE corpus, we use the same split into training and test data as in Stab and Gurevych (2017). In particular, our test data comprises 80 documents ("essays") with a total of 29,537 tokens (en). We choose 10% of the training data as dev set. Thus, we have 286 essays in the train set with a total of 105,988 tokens (en) and 36 essays in the dev set with a total of 12,657 tokens (en). We report averages over five random initializations of our networks. For the CRC corpus, we perform 5-fold cross-validation on document level. Our train sets consist of roughly 15K tokens, our dev sets of 2K tokens, and our test sets of 4K tokens. For each split, we average over five different random initializations and report the average over these averages. For MTX, we also perform 5-fold cross-validation on document level. Our train sets consist of roughly 6K tokens (en), our dev sets of 500 tokens (en), and our test sets of 1,500 tokens (en). We use the same averaging strategy as for CRC, but average over ten random initializations per fold, to account for the smaller dataset sizes.

5.2 Results

We report results for adapting between datasets in different languages and between parallel versions of one and the same dataset. We only consider cases where one of the involved languages is English. Further, we do not transfer between MTX and the other datasets, because MTX has no "O" units (and no major claims). Unless stated otherwise, we always *evaluate* on `HT` for both direct transfer and projection.

5.2.1 Direct Transfer

For all cross-lingual direct transfer experiments, we train on the union of train and dev (train+dev) sets (randomly drawn for the datasets for which we used cross-validation) of the source language and test either on the whole data (train+dev+test) of L2, or, in case of parallel versions of a dataset (such as PE EN↔DE) on the test set of L2. We do not use `MT` for direct transfer at any stage.

PE_{EN}↔PE_{DE}, results in Table 4: English in-language results do not vary much and are on a level of slightly above 69% macro-F1, largely independent of the embedding types and whether or not character

information is available.[4] German in-language results are 4-5 percentage points (pp) below the English ones. One might suspect the presumed inferior quality (or derivative nature) of the student translations as a cause for this, but we hypothesize that German is simply more complex than English, both in morphology and syntax.

We observe a noticeable drop when moving cross-language. This drop is up to >40% for the direction EN→DE (worst case drop from ∼70% to ∼37%) and slightly less for DE→EN. We explain this drop by the discrepancy between training and test distributions. This discrepancy is present even in bilingual embedding spaces: no test word has the exact same representation as the words in the training data. Further, disabling character information typically has a very positive impact cross-language. For example, EN→DE performance increases from ∼40% F1 to ∼50% when disabling character information. The reason is that a system that extracts a character representation based on English characters may get confused from the diverging German character sequences. Surprisingly, characters do not impede so much in the direction DE→EN. The reason seems to be lexical borrowing in modern German from English. For example, ∼17% of the 'active' vocabulary (i.e., frequency >30) of English in PE_{EN} is also contained in PE_{DE}. In contrast, only 6% of the active vocabulary of German in PE_{DE} occurs also in PE_{EN}.

CRC↔PE_{EN}, results in Table 5 (left): In-language CRC results are lower than in-language PE results (∼46% vs. ∼69% for PE). This is unsurprising since CRC is considerably smaller in size than PE. However, we observe that the cross-language drop is much larger than it is for the PE DE↔EN setting. In fact, performance values always lie below our random baseline. We attribute this huge drop not to the larger language distance between English and Chinese (relative to English and German), but primarily to the domain gap between student essays and hotel reviews. In fact, we observe that, e.g., major claims in PE are almost always preceded by specific discourse markers such as "Therefore, I believe that" or "In the end, I think" (cf. Table 2), while hotel reviews completely lack such discourse connectives (cf. Table 3). Since we expect a system that trains on PE to learn the signaling value of these markers, directly applying this system to text where such markers are absent, fails.

MTX_{EN}↔MTX_{DE}, results in Table 5 (right): Even though the dataset is by far smallest in size it yields the highest F1-scores among all our considered datasets. Moreover, the language drop is comparatively small (between 4 and 7pp). Investigating, we notice that argument components are typically separated by punctuation symbols (mostly "." or ",") in MTX, which is easy to learn even cross-lingually. Moreover, we find that claims can often be separated from premises by simple keywords such as "should", which can, apparently, be reliably spotted cross-lingually via the corresponding bilingual word embeddings.

Model	Embedding Type	EN→EN	EN→DE	DE→DE	DE→EN
BLCRF+Char	BIVCD-100	68.87	41.89	65.22	49.91
	BIVCD-200	**70.51**	39.87	**65.92**	49.52
	BISKIP-100	69.27	37.01	63.33	48.23
BLCRF	BIVCD-100	69.27	49.70	65.90	50.14
	BISKIP-100	69.15	**49.76**	64.92	**50.28**
Baseline		20.	20.	20.	20.

Table 4: Direct transfer results for PE_{EN}↔PE_{DE}. Scores are macro-F1.

Error Analysis and Discussion For PE direct transfer experiments, we find that a major source of errors is incorrect classification of tokens labeled "B-". This means that the system has difficulty finding the exact beginning of an argument span. We find, however, that the reason for the language drop is not that the bilingual embedding spaces are bad: among the top-10 neighbors of English words are roughly five German words, and vice versa. Rather, direct transfer induces a situation very similar to standard monolingual out-of-vocabulary (OOV) scenarios, namely as if all test words had been replaced by synonyms that did not occur in the train data. While systems using embeddings as input are more robust to

[4]Our in-language results are slightly below our previous results reported in Eger et al. (2017) (table 6), where we obtained scores of 72-75% for token-level component extraction, even though the architecture is in principle the same. Reasons may be the different word embeddings used as well as that we reported majority performance over different hyperparameter combinations in the previous work, which typically increased performance scores by a few percentage points.

Model	CRC↔PE_{EN}				MTX_{EN}↔MTX_{DE}			
	ZH→ZH	ZH→EN	EN→EN	EN→ZH	EN→EN	EN→DE	DE→DE	DE→EN
BLCRF+Char	**46.31**	14.01	**69.27**	9.50	**73.12**	67.03	**73.41**	**66.62**
BLCRF	44.95	16.52	69.15	12.60	72.15	**69.46**	72.52	63.71
Baseline	18.	**17.**	20.	**17.**	45.	46.	50.	50.

Table 5: Direct transfer results for CRC and MTX. Scores are macro-F1. Embeddings are BISKIP-100.

OOV words, they are still affected by them (Ma and Hovy, 2016; Müller et al., 2013). This "blurring effect" at test time then makes it more difficult to detect exact argument component spans. While this is true in general, it is not true for punctuation symbols, which typically have an identical role across languages and, hence, extremely similar representations. For example, German and English "." have more than 97% cosine similarity in BISKIP-100d, which is much higher than for typical monolingually closely related words. Finally, besides semantic shift direct transfer also faces syntactic shift, because the test words may have different word order compared to the train data (e.g., verb final position in German).

The lessons we learn from our above experiments are that (i) the MTX dataset does not provide a real challenge for cross-lingual techniques because argument components can easily be spotted based on punctuation and component typing appears to be just as easily portable across languages. (ii) Language adaptation between the CRC corpus and PE appears, in contrast, too difficult because argumentation units are very differently realized across the two datasets, and hence, the domain shift appears to be the (much) larger obstacle compared to the language shift.[5] Thus, (iii) we mostly focus on the cross-lingual version of the PE corpus in the sequel, which is a difficult enough dataset for cross-lingual AM, without confounding the problem with issues relating to differences of AM domains. In addition, we only use BISKIP-100d embeddings, because the choice of embeddings seemed to have a negligible effect in our case (certainly, for our main focus, namely, cross-lingual evaluations) and because they showed slightly superior results cross-lingually than BIVCD-100d.

5.2.2 Projection

HT-Projection Table 6 (HT columns) shows results when we project PE_{EN} training data annotations on parallel German HT documents and train and evaluate a system directly on German (and vice versa). As said before, we train in this case directly on the same language as we test on, viz., L2. We observe that this improves cross-language results dramatically. From a best cross-language result of 49.76% for BIVEC-100d in the direct transfer setting, we improve by almost 30% to 63.67%. This is only roughly 1pp below the best in-language result for German which was 64.92%. In the direction DE→EN, we observe the same trend: we improve by over 30% relative to the direct transfer results and achieve a macro-F1 score that is only 1.7pp below the in-language upper-bound.

MT-Projection Next, we investigate what happens when we replace the HT translations with MT translations and perform the same projection steps as before. Results for PE_{EN}↔PE_{DE} are shown in Table 6 (MT columns). We see that EN→DE results get slightly better while DE→EN results get slightly worse. On average, it seems, using machine translations is just as good as using human translations. Moreover, we remain very close to the upper-bound in-language results. The reason why MT results could be better than HT results is that the machine might translate more consistently. It might also be better in certain cases in correcting (the sometimes ungrammatical) English original. Another likely reason is that current MT has reached, if not already surpassed, HT of non-expert (but bilingually fluent) human translators.[6]

Motivated by this finding, we also machine translated the CRC corpus into English, projected annotations and then trained a system on this translation and evaluated on the PE_{EN} corpus. Results improve to 23.15% macro-F1 score relative to the best direct transfer result of 16.52%. This indicates, again, that

[5] This is a similar finding as in Daxenberger et al. (2017).

[6] We conducted a formal test if MT can reliably be distinguished from HT in our setup. We trained a system (an adaptation of InferSent (Conneau et al., 2017)) to predict whether for an English original e and a second input sentence it could determine if the second input is a human or machine translation of e. The system's performance of 54% accuracy (which is only slightly better than random guessing) matched our own intuition and introspection into the quality of the machine translations.

training in L2 is better than training in L1, given the high quality machine translations and a suitable projection algorithm. However, 23.15% is still only slightly better than the random baseline of 17%—and far from the best PE_{EN} in-language result of >69%. In addition, in the direction PE_{EN}→CRC macro-F1 (also) improves to 15.33% relative to a best direct transfer score of 12.60%. However, here the performance is still below the random baseline of 17%. We take this as strong evidence that the domain gap between CRC and PE is too large and it is not possible to train a system in one of these two domains and directly apply it in the other, even when the language gap has been eliminated.

	EN→DE			DE→EN		
	HT	MT	In-Lang.	HT	MT	In-Lang.
BLCRF+Char	63.67	**64.00**	63.33	**67.57**	66.39	69.27
BLCRF	61.18	63.34	64.92	64.87	64.68	69.15

Table 6: Projection on HT/MT translations, evaluated on human-created test data. Scores are macro-F1. Embeddings are BISKIP-100.

Other languages We conducted a final experiment in which we considered our MT translations of PE into French, Spanish, and Chinese. Since we have no human-created test data for these languages we could only evaluate on *machine translations and projected annotations*. For our BLCRF+char model, we obtained performance scores of 62.45%, 65.92%, 59.20% for French, Spanish and Chinese, respectively. To see if these numbers give reliable estimates of the systems when evaluated on HT data, we performed the same test with German and English and got scores of 63.20% and 61.45%, respectively, when trained and evaluated on MT. For CRC, we also trained and evaluated on the English (MT translated and automatically projected) data, obtaining a score of 47.92% with BLCRF+char, which is slightly above the in-language value of 46.31% on the original Chinese data (see Table 5). That all these numbers are close to the original in-language results gives a good indication that the MT evaluations very likely strongly correlate to 'true' performances on HT data.

Error analysis The bottleneck of projection is the quality of the cross-lingual projections (which in turn depend on the quality of the word alignments between bi-text). We can directly assess our projections by comparing them to the human-created annotations. Our algorithmic projections match in 97.24% of the cases (token-level) with the human gold standard for the direction EN→DE. The corresponding macro-F1 score is 89.85%. Inspecting the confusion matrices, we observe that most mismatches occur between the "B" and "I" categories of a given component type and with the "O" category. These numbers and the mismatch types indicate that there are only few projection errors and they typically lead to either (slightly) larger or smaller argument components than given in the human-created data. To illustrate, typical projection errors arise in case of missing articles in one of the two languages involved; cf. Table 2: "luck is [...]" vs. "la chance est [...]". Here, it is likely that the alignment algorithm does not align the French determiner "la" to an English word, and thus, "la" is not included in the argument component. Another case of too short argument components is that of verb final position in German which often gets misaligned and the corresponding final verb omitted from the argument component. These misalignments lead to slightly "shifted" argument components in the L2 train set and are the most prominent source of error for the projection technique. To quantify: the German in-language system classifies 58% of all cases when one of the determiners ("der/die/das") begins an argument component correctly, but the system using projections from the English data only classifies 35% of them correctly. Hence, alignment/projection errors indeed propagate to some degree, but these phenomena are rare and have negligible impact on performance. Note that, by design of our projection algorithm, misalignments of words in the 'center' of an argument component are much less likely to be a problem.

Figure 1 plots individual F1 scores for various systems transferring to English on PE. Here, the cross-lingual systems using HT/MT projection perform roughly as well as the in-language system for "O", "I-C" and "I-P". These are the most frequent classes. For claims and major claims, which have lower frequency, the in-language upper bound tends to perform better. Noteworthy, the in-language system is

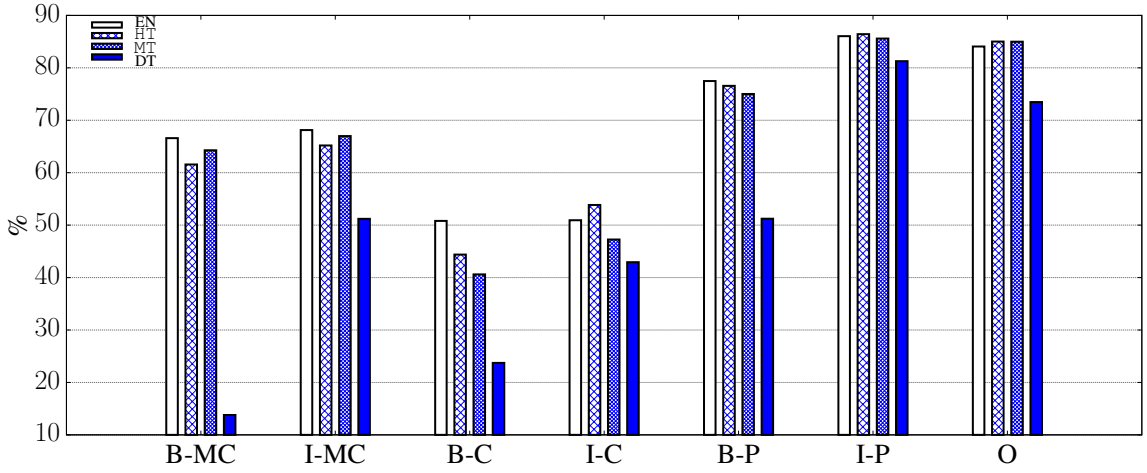

Figure 1: Individual F1-scores for four indicated systems and seven labels. All transfer systems are from PE_{DE} to PE_{EN}; EN is in-language. DT stands for Direct Transfer. HT/MT are projection-based approaches. Embeddings are BISKIP-100. Systems are BLCRF+Char.

always better for the beginnings of an argument component (B-MC,B-C,B-P), which confirms our above analysis. The direct transfer system, in contrast, performs much worse, particularly for major claims and claims, and also for all starts of components, indicating that the blurring ("OOV") effect is here much more severe.

6 Conclusion

Showing that the currently available datasets for AM are not adequate for evaluating cross-lingual AM transfer, we created human and machine translations of one of the most popular current AM datasets, the dataset of persuasive student essays (Stab and Gurevych, 2017). We also machine translated a Chinese corpus of reviews (Li et al., 2017) into English, which provides argumentation structures on hotel reviews. Performing cross-lingual experiments using suitable adaptations of two popular transfer approaches, we have shown that machine translation and (naive) projection work considerably better than direct transfer, even though the former approach contains two sources of noise. Moreover, machine translation in combination with projection almost performs on the level of in-language upper bound results. We think that our findings shed further light on the value—and the huge potential—of current (neural) machine translation systems for cross-lingual transfer. They also cast doubt on current standard use of direct transfer in cross-lingual scenarios. Instead, we propose to simply machine translate the train set, when this is possible, and then project labels to the translated text. This eliminates the (particular) "OOV" and "ordering" problems inherent to direct transfer. Prerequiste to this approach is high quality MT, which, with the advent of neural techniques, appears to be now available.

We hope our new datasets fuel AM research in languages other than English. In this work, we did not consider cross-lingual argumentative relation identification, although relations are available in the newly created parallel PE and CRC datasets. Future work should explore cross-lingual multi-task learning for AM (Schulz et al., 2018) with the source language as main task and small amounts of labeled target language data, as well as adversarial training techniques (Yasunaga et al., 2018), which promise to be beneficial for the particular OOV problem that direct transfer is prone to (though not for the ordering problem). We also want to combine projection with direct transfer by training on the union of projected L2 data as well as the original L1 data using shared representations.

Acknowledgements

This work has been supported by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1816B (CEDIFOR) and 03VP02540 (ArgumenText).

References

Zeljko Agic, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *TACL*, 4:301–312.

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ahmet Aker and Huangpan Zhang. 2017. Projection of argumentative corpora from source to target languages. In *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 67–72.

Pierpaolo Basile, Valerio Basile, Elena Cabrio, and Serena Villata. 2016. Argument mining on italian news blogs. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017. Cross-lingual and cross-domain discourse segmentation of entire documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 237–243.

Xilun Chen, Ben Athiwaratkun, Yu Sun, Kilian Q. Weinberger, and Claire Cardie. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *Arxiv preprint https://arxiv.org/abs/1612.08994*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, pages 681–691. Association for Computational Linguistics.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 759–770. Association for Computational Linguistics.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 600–609.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the Essence of a Claim? Cross-Domain Claim Identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 2055–2066, September.

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

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242, Lisbon, Portugal.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada, July. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2822–2828. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Patrik Lambert. 2015. Aspect-level cross-lingual sentiment classification with constrained smt. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 781–787, Beijing, China, July. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Mengxue Li, Shiqiang Geng, Yang Gao, Shuhua Peng, Haijing Liu, and Hao Wang. 2017. Crowdsourcing Argumentation Structures in Chinese Hotel Reviews. In *Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics*, pages 87–92, Banff, Canada.

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld. In *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. 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. The Association for Computational Linguistics.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

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*.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 62–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Arxiv preprint https://arxiv.org/abs/1301.3781*.

Raquel Mochales-Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107, Barcelona, Spain.

Marie-Francine Moens. 2017. Argumentation mining: How can a machine acquire common sense and world knowledge? *Argument & Computation*. Accepted.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA, October. Association for Computational Linguistics.

Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137, Berlin, Germany, August. Association for Computational Linguistics.

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

Andreas Peldszus and Manfred Stede. 2013a. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31.

Andreas Peldszus and Manfred Stede. 2013b. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204, Sofia, Bulgaria.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation*, pages 801–815, Lisbon, Portugal.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulic, and Anders Sogaard. 2017. A survey of cross-lingual word embedding models. In *Arxiv preprint https://arxiv.org/abs/1706.04902*.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66, Denver, CO.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–41. Association for Computational Linguistics, June.

Alfred Sliwa, Yuan Man, Ruishen Liu, Niravkumar Borad, Seyedeh Ziyaei, Mina Ghobadi, Firas Sabbah, and Ahmet Aker. 2018. Multi-lingual argumentative corpora in english, turkish, greek, albanian, croatian, serbian, macedonian, bulgarian, romanian and arabic. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, September.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 477–487, Stroudsburg, PA, USA. Association for Computational Linguistics.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan T. McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *TACL*, 1:1–12.

Chen-Tse Tsai, Stephen D. Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016*, pages 219–228.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. In *Fourth SIGHAN Workshop on Chinese Language Processing*.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Ivan Vulic and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *ACL (2)*, pages 719–725. The Association for Computer Linguistics.

Henning Wachsmuth, Giovanni Da San Martino, Dora Kiesel, and Benno Stein. 2017. The Impact of Modeling Overall Argumentation with Tree Kernels. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 17)*, pages 2369–2379. Association for Computational Linguistics, September.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *Arxiv preprint https://arxiv.org/abs/1703.06345*.

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

Michihiro Yasunaga, Jungo Kasai, and Dragomir R. Radev. 2018. Robust multilingual part-of-speech tagging via adversarial training. In *Proceedings of NAACL*. Association for Computational Linguistics.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi S. Jaakkola. 2016. Ten pairs to tag - multilingual POS tagging via coarse mapping between embeddings. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1307–1317.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, page 1403–1412.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may.