# Annotation Projection-based Representation Learning for Cross-lingual Dependency Parsing

**Min Xiao** and **Yuhong Guo**
Department of Computer and Information Sciences
Temple University, Philadelphia, PA 19122, USA
{minxiao,yuhong}@temple.edu

## Abstract

Cross-lingual dependency parsing aims to train a dependency parser for an annotation-scarce target language by exploiting annotated training data from an annotation-rich source language, which is of great importance in the field of natural language processing. In this paper, we propose to address cross-lingual dependency parsing by inducing latent cross-lingual data representations via matrix completion and annotation projections on a large amount of unlabeled parallel sentences. To evaluate the proposed learning technique, we conduct experiments on a set of cross-lingual dependency parsing tasks with nine different languages. The experimental results demonstrate the efficacy of the proposed learning method for cross-lingual dependency parsing.

## 1 Introduction

The natural language processing (NLP) community has witnessed an enormous development of multilingual resources, which draws increasing attention to developing cross-lingual NLP adaptation systems. Cross-lingual dependency parsing aims to train a dependency parser for a target language where labeled data is rare or unavailable by exploiting the abundant annotated data from a source language. Cross-lingual dependency parsing can effectively reduce the expensive manual annotation effort in individual languages and has been increasingly studied in the multilingual community. Previous works have demonstrated the success of cross-lingual dependency parsing for a variety of languages (Durrett et al., 2012; McDonald et al., 2013; Täckström et al., 2013; Søgaard and Wulff, 2012).

One fundamental issue of cross-lingual dependency parsing lies in how to effectively transfer the *annotation* information from the source language domain to the target language domain. Due to the language divergence over the word-level representations and the sentence structures, simply training a monolingual dependency parser on the labeled source language data without adaptation learning will fail to produce a dependency parser that works in the target language domain. To tackle this problem, a variety of works in the literature have designed better algorithms to exploit the annotated resources in the source languages, including the cross-lingual annotation projection methods (Hwa et al., 2005; Smith and Eisner, 2009; Zhao et al., 2009), the cross-lingual direct transfer with linguistic constraints methods (Ganchev et al., 2009; Naseem et al., 2010; Naseem et al., 2012), and the cross-lingual representation learning methods (Durrett et al., 2012; Täckström et al., 2012; Zhang et al., 2012).

In this work, we propose a novel representation learning method to address cross-lingual dependency parsing, which exploits annotation projections on a large amount of unlabeled parallel sentences to induce latent cross-lingual features via matrix completion. It combines the advantages of the cross-lingual annotation projection methods, which project labeled information into the target language domain, and the cross-lingual representation learning methods, which learn latent interlingual features. Specifically, we first train a dependency parser on the labeled source language data and use it to infer labels for the unlabeled source language sentences of the parallel resources. We then project the annotations from the source language to the target language via the word alignments on the parallel sentences. Afterwards, we define a set of interlingual features and construct a word-*feature* matrix by associating each word with these language-independent features. We then use the original labeled source language data and the predicted (or projected) labeled information on the parallel sentences to fill in the observed entries of the word-feature matrix, while matrix completion is performed to fill the remaining missing entries. The completed word-feature matrix provides a set of consistent cross-lingual representation features for the words in both languages. We use these features as augmenting features to train a dependency parsing system on the labeled data in the source language and perform prediction on the test sentences in the target language. To evaluate the proposed learning method, we conduct experiments on eight cross-lingual dependency parsing tasks with nine different languages. The experimental results demonstrate the superior performance of the proposed cross-lingual transfer learning method, comparing to other approaches.

## 2 Related Work

A variety of cross-lingual dependency parsing methods have been developed in the literature. We provide a brief review over the related works in this section.

Much work developed in the literature is based on annotation projection (Hwa et al., 2005; Liu et al., 2013; Smith and Eisner, 2009; Zhao et al., 2009). Basically, they exploit parallel sentences and first project the annotations of the source language sentences to the corresponding target language sentences via the word level alignments. Then, they train a dependency parser in the target language by using the target language sentences with projected annotations. The performance of annotation projection-based methods can be affected by the quality of word-level alignments and the specific projection schema. Therefore, Hwa et al. (2005) proposed to heuristically correct or modify the projected annotations in order to increase the projection performance while Smith and Eisner (2009) used a more robust projection method, quasi-synchronous grammar projection, to address cross-lingual dependency parsing. Moreover, Liu et al. (2013) proposed to project the discrete dependency arcs instead of the treebank as the training set. These works however assume that the parallel sentences are already available, or can be obtained by using free machine translation tools. Instead, Zhao et al. (2009) considered the cost of machine translation and used a bilingual lexicon to obtain a translated treebank with projected annotations from the source language.

A number of works are developed based on representation learning (Durrett et al., 2012; Täckström et al., 2012; Zhang et al., 2012; Xiao and Guo, 2014). In general, these methods first automatically learn some language-independent features and then train a dependency parser in this interlingual feature space with labeled data in the source language and apply it on the data in the target language. Durrett et al. (2012) used a bilingual lexicon, which can be manually constructed or induced on parallel sentences, to learn language-independent projection features for cross-lingual dependency parsing. Täckström et al. (2012) used unlabeled parallel sentences to induce cross-lingual word clusterings and used these word clusterings as interlingual features. Both (Durrett et al., 2012) and (Täckström et al., 2012) assumed that the twelve universal part-of-speech (POS) tags (Petrov et al., 2012) are available and used them as the basic interlingual features. Moreover, Zhang et al. (2012) proposed to automatically map language-specific POS tags to universal POS tags to address cross-lingual dependency parsing, instead of using the manually defined mapping rules. Recently, Xiao and Guo (2014) used a set of bilingual word pairs as pivots to learn interlingual distributed word representations via deep neural networks as augmenting features for cross-lingual dependency parsing.

Some other works are proposed based on multilingual linguistic constraints (Ganchev et al., 2009; Gillenwater et al., 2010; Naseem et al., 2010; Naseem et al., 2012). Basically, they first construct a set of linguistic constrains and then train a dependency parsing system by incorporating the linguistic constraints via posterior regularization. The constraints are expected to bridge the language differences. Ganchev et al. (2009) automatically learned the constraints by using parallel data while some other works manually constructed them by using the universal dependency rules (Naseem et al., 2010) or the typological features (Naseem et al., 2012).

## 3 Proposed Approach

In this section, we present a novel representation learning method for cross-lingual dependency parsing, which combines annotation projection and matrix completion-based feature representation learning together to produce effective interlingual features.

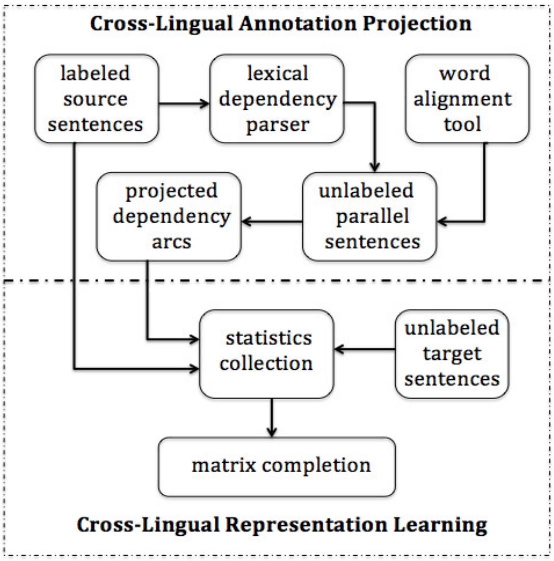
Cross-Lingual Annotation Projection
labeled source sentences
lexical dependency parser
word alignment tool
projected dependency arcs
unlabeled parallel sentences
statistics collection
unlabeled target sentences
matrix completion
Cross-Lingual Representation Learning

Figure 1: The architecture of the proposed cross-lingual representation learning framework, which consists of two steps, cross-lingual annotation projection and cross-lingual representation learning.

We consider the following cross-lingual dependency parsing setting. We have a large amount of labeled sentences in the source language and a set of unlabeled sentences in the target language. In addition, we also have a large set of auxiliary unlabeled parallel sentences across the two languages. We aim to learn interlingual feature representations such that a dependency parser trained in the source language sentences can be applied in the target language domain. The framework for the proposed cross-lingual representation learning system is given in Figure 1. The system has two steps: cross-lingual annotation projection and cross-lingual representation learning. We present each of the two steps below.

### 3.1 Cross-Lingual Annotation Projection

In the first step, we employ a large amount of unlabeled parallel sentences to transfer dependency relations from the source language to the target language. We first train a lexicalized dependency parser with the labeled training data in the source language. Then we use this parser to produce parse trees on the source language sentences of the auxiliary parallel data. Simultaneously, we perform word-level alignments on the unlabeled parallel sentences using existing alignment tools. Finally, we project the predicted dependency relations of the source language sentences to their parallel counterparts in the target language via the word-level alignments. Instead of projecting the whole dependency trees, which requires more sophisticated algorithms, we simply project each dependency arc on the source sentences to the target language side.

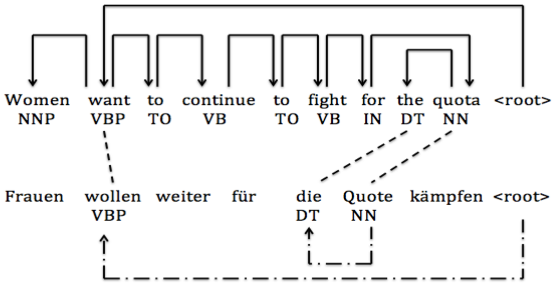

Figure 2: An example of cross-lingual annotation projection, where a partial word-level alignment is shown to demonstrate two cases of annotation projection.

We now use a specific example in Figure 2 to illustrate the projection step. This example contains an English sentence and its parallel sentence in German. The English sentence is fully labeled with each dependency relation indicated by a solid directed arc. The dashed lines between the English sentence and the German sentence show the alignments between them. For each dependency arc instance, we consider the following properties: the parent word, the child word, the parent POS, the child POS, the dependency direction, and the dependency distances. The projection of the dependency relations from the source language to the target language is conducted based on the word-level alignment. There are two different scenarios. The first scenario is that the two source language words involved in the dependency relation are aligned to two different words in the corresponding target sentence. For example, the English words “the” and “quota” are aligned to German words “die” and “Quote” separately. We then copy this dependency relation into the target language side. The second scenario is that a source language word is aligned to a word in the target language sentence and has a dependency relation with the “<root>” word. For example, the English word “want” is aligned to “wollen” and it has a dependency arc with “<root>”. We then project the dependency relation from the English side to the German side as well. Moreover, we also directly project the POS tags of the source language

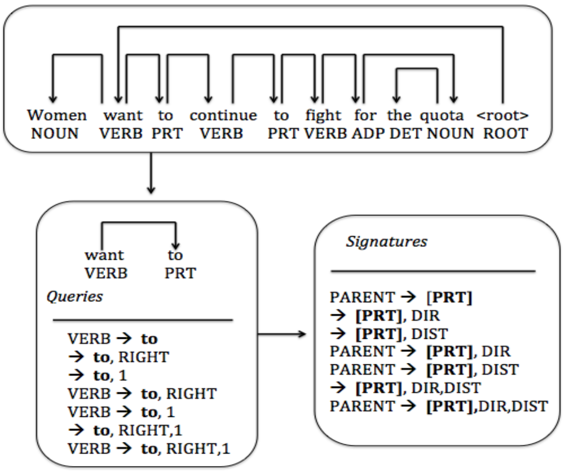

Figure 3: Example of how to collect queries for each specific dependency relation and how to obtain the abstract signatures (adapted from (Durrett et al., 2012)).

words onto the target language words. Since the word order for each aligned word pair in parallel sentences can be different, we recalculate the dependency direction and the dependency distance for the projected dependency arc instance. Note the example in Figure 2 only shows a partial word-level alignment to demonstrate the two cases of the annotation projection. The word alignment tool can align more words than shown in the example.

## 3.2 Cross-Lingual Representation Learning

After cross-lingual annotation projection, we have a set of projected dependency arc instances in the target language. However, the sentences in the target language are not fully labeled. Dependency relation related features are not readily available for all the words in the target language domain. Hence, in this step, we first generate a set of interlingual features and then automatically fill the missing feature values for the target language words with matrix completion based on the projected feature values.

### 3.2.1 Generating Interlingual Features

We use the signature method in (Durrett et al., 2012) to construct a set of interlingual *features* for the words in the source and target language domains . The signatures proposed in (Durrett et al., 2012) for dependency parsing are universal across different languages, and have numerical values that are computed in specific dependency relations. Here we illustrate the signature generation process by using an example in Figure 3, which is adapted from (Durrett et al., 2012). Note for each dependency relation between a parent (also known as the head) word and a child (also known as the dependent) word, we can collect a number of queries based on the dependency properties. For example, given the dependency arc between "want" and "to" in the English sentence in Figure 3, and assuming we consider the child word "to", we produce queries by considering a non-empty subset of the dependency properties (the parent POS, the dependency direction, the dependency distance), which provides us 7 queries: "VERB→to", "→to, RIGHT", "→to, 1", "VERB →to, RIGHT", "VERB→to, 1", "→to, RIGHT, 1", "VERB→to, RIGHT, 1", where VERB is the parent POS tag, RIGHT is the dependency direction and 1 is the dependency distance. Then we can abstract the specific queries to generate the signatures by replacing the considered word ("to") with its POS tag ("PRT"), and replacing the parent POS tag with "PARENT", the specific dependency distance with "DIST" and the dependency direction with "DIR". This produces the following 7 signatures: "PARENT→[PRT]", "→[PRT], DIR", "→[PRT], DIST", "PARENT→[PRT], DIST", "PARENT→[PRT], DIST", "→[PRT], DIR, DIST", and "PARENT→[PRT], DIR, DIST", where the brackets indicate the POS tags are for the considered word. Similarly, we can perform the same abstraction process for the parent word "want" and get another 7 signatures (see Table 1). Since each signature contains one POS tag and there are 13 different POS types (12 universal POS tags and 1 special type for the "<root>" word), we can get a total of 7 × 2 × 13 = 182 signatures. These signatures are independent of specific languages, though their numerical values should be computed in a specific dependency relation for each considered target word.

A set of interlingual *features* can then be generated from these abstractive signatures by considering different instantiations of their items. For a given target word with an observed POS tag, it has 14 signatures (see Table 1). For each signature, we consider all possible instantiations of its other items given the fixed target word. For example, for the target word "to", its signature "→[PRT], DIR" can be instantiated into 2 features: "→ LEFT" and "→ RIGHT". Similarly, its signature "→[PRT],

| Signatures | # Features |
|---|---|
| [PRT] → DIR | 2 |
| [PRT] → DIST | 5 |
| [PRT] → CHILD | 13 |
| [PRT] → DIR, DIST | 10 |
| [PRT] → CHILD, DIR | 26 |
| [PRT] → CHILD, DIST | 65 |
| [PRT] → CHILD, DIR, DIST | 130 |
| → [PRT], DIR | 2 |
| → [PRT], DIST | 5 |
| PARENT → [PRT] | 13 |
| → [PRT], DIR, DIST | 10 |
| PARENT → [PRT], DIR | 26 |
| PARENT → [PRT], DIST | 65 |
| PARENT → [PRT], DIR, DIST | 130 |
| Total | 502 |

Table 1: The number of induced "features" of each signature for a given word.

DIST" can be instantiated into 5 features since DIST has 5 different values ({1, 2, 3–5, 6–10, 11+}), and its signature "[PRT]→CHILD" can be instantiated into 13 features since CHILD denotes the child word's POS tags and can have 13 different values. Hence as shown in Table 1, we can get 502 features from the 14 signatures.

### 3.2.2 Learning Feature Values with Matrix Completion

The signature-based 502 interlingual features together with the 13 universal POS tag features can be used as *language independent features* for all the words in the vocabulary constructed across the source and target language domains. In particular, we can form a *word-feature* matrix with the constructed vocabulary and the total 515 language independent features. For each word that appeared in the dependency relation arcs, we can use the number of appearances of its interlingual features as the corresponding feature values. However, the sentences in the target language are not fully labeled. Some words in the target language domain may not be observed in the projected dependency arc instances, and we cannot compute their feature values for the 502 interlingual features, though the 13 universal POS tag features are available for all words. Moreover, since we only have a limited number of projected dependency arc instances in the target language, even for some target words that appeared in the projected arc instances of the parallel data, we may only observe a subset of features among the total 502 interlingual features, with the rest features missing. Hence the constructed *word-feature* matrix is only partially observed, as shown in Figure 4. Furthermore, there could also be some noise in the observed feature values as some word features may not have received sufficient observations.

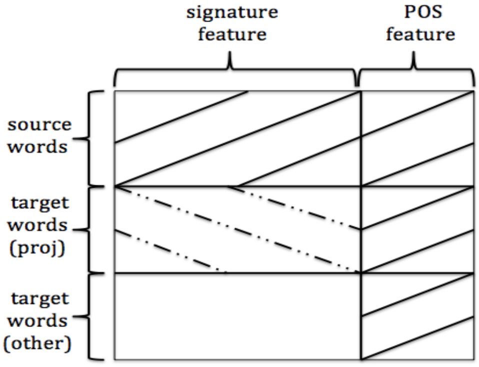

Figure 4: The word-feature matrix. There are three parts of words: the source language words, target language words from the projected dependency arc instances, and additional target language words. The signature features are the 502 interlingual features and the POS features are the 13 universal POS tags. Solid lines indicate observed entries, dashed lines indicate partially observed entries, while empty indicates missing entries.

To solve the missing feature problem and simultaneously perform data denoising, we exploit a feature correlation assumption: the 502 constructed interlingual features and the 13 universal POS tags are not mutually independent; they usually contain a lot statistical correlation information. For example, for a word "want" with POS tag "VERB", its feature value for "VERB → **want**, RIGHT" is likely to be very small such as zero, while its feature value for "**want**→ NOUN, LEFT" is likely to be large. Moreover, the existence of any one of the two interlingual features in this example can also indicate the non-existence of the other feature. The existence of feature correlations establishes the low-rank property of the word-feature matrix. We hence propose to fill the missing feature values and reduce the noise in the word-feature matrix by performing matrix completion. Low-rank matrix completion has been successfully used in many applications to fill missing entries of partially observed low-rank matrices and perform matrix denoising (Cabral et al., 2011; Xiao and Guo, 2013) by exploiting the feature correlations and underlying low-dimensional representations. Following the same principle, we expect to automatically discover the missing feature values in our word-feature matrix and perform denoising through low-rank matrix completion.

Let $M^0 \in \mathbb{R}^{n \times k}$ denote the partially observed word-feature matrix in Figure 4, where $n$ is the number of words and $k$ is the dimensionality of the feature set, which is 515 in this study. Let $\Omega$ denote the set of indices for the observed entries. Hence for each observed entry $(i, j) \in \Omega$, $M^0_{ij}$ contains the frequency collected for the $j$-th feature of the $i$-th word. We then formulate matrix completion as the following optimization problem to recover a full matrix $M$ from the partially observed matrix $M^0$:

$$\min_{M \geq 0} \gamma \|M\|_* + \alpha \|M\|_{1,1} + \sum_{(i,j) \in \Omega} (M_{ij} - M^0_{ij})^2 \quad (1)$$

where the trace norm $\|M\|_*$ enforces the low-rank property of the matrix, and $\|M\|_{1,1}$ denotes the entrywise L1 norm. Since many words usually only have observed values for a small subset of the 502 interlingual features due to the simple fact that they are only associated with very few POS tags, a fully observed word-feature matrix is typically sparse and contains many zero entries. Hence we use the L1 norm regularizer to encode the sparsity of the matrix $M$. The nonnegativity constraint $M \geq 0$ encodes the fact that our frequency based feature values in the word-feature matrix are all nonnegative. The minimization problem in Eq (1) can be solved using a standard projected gradient descent algorithm (Xiao and Guo, 2013).

### 3.3 Cross-Lingual Dependency Parsing

After matrix completion, we can get a set of interlingual features for all the words in the word-feature matrix. We then use the interlingual features for each word as augmenting features and train a delexicalized dependency parser on the labeled sentences in the source language. The parser is then applied to perform prediction on the test sentences in the target language, which are also delexicalized and augmented with the interlingual features.

## 4 Experiments

### 4.1 Datasets

We used the multilingual dependency parsing dataset from the CoNLL-X shared tasks (Buchholz and Marsi, 2006; Nivre et al., 2007) and experimented with nine different languages: *Danish (Da), Dutch (Nl), English (En), German (De), Greek (El), Italian (It), Portuguese (Pt), Spanish (Es)* and *Swedish (Sv)*. For each language, the original dataset contains a training set and a test set. We constructed eight cross-lingual dependency parsing tasks, by using English as the label-rich source language and using each of the other eight languages as the label-poor target language. For example, the task *En2Da* means that we used English sentences as the source language data and Danish sentences as the target language data. For each task, we used the original training set in English as the labeled source language data, and used the original training set in the target language as unlabeled training data and the original test set in the target language as test sentences. Each sentence from the dataset is labeled with gold standard POS tags. We manually mapped these language-specific POS tags to 12 universal POS tags: NOUN (nouns), NUM (numerals), PRON (pronouns), ADJ (adjectives), ADP (prepositions or postpositions), ADV (adverbs), CONJ (conjunctions), DET (determiners), PRT (particles), PUNC (punctuation marks), VERB (verbs) and X (for others).

We used the unlabeled parallel sentences from the *European parliament proceedings parallel corpus* (Koehn, 2005), which contains parallel sentences between multiple languages, as auxiliary unlabeled parallel sentences in our experiments. For the representation learning over each cross-lingual dependency parsing task, we used all the parallel sentences for the given language pair from this corpus. The number of parallel sentences for the eight language pairs ranges from $1,235,976$ to $1,997,775$, and the number of tokens involved in these sentences in each language ranges from $31,929,703$ to $50,602,994$.

### 4.2 Representation Learning

For the proposed representation learning, we first trained a lexicalized dependency parser on the labeled source language data using the MSTParser tool (proj with the first order set) (McDonald et al., 2005) and used it to predict the parsing annotations of the source language sentences in the unlabeled parallel dataset. The sentences of the parallel data only contain sequences of words, without additional POS tag information. We then used an existing POS tagging tool (Collobert et al., 2011) to infer POS tags for them. Next we produced word-level alignments on the unlabeled parallel

| Basic | Conj with dist | Conj with dir | Conj with dist and dir |
|---|---|---|---|
| upos_h | dist, upos_h | dir, upos_h | dist, dir, upos_h |
| upos_d | dist, upos_d | dir, upos_d | dist, dir, upos_d |
| upos_h, upos_d, | dist, upos_h, upos_d | dir, upos_h, upos_d | dist, dir, upos_h, upos_d |

Table 2: Feature templates for training a basic delexicalized dependency parser. *upos* stands for the universal POS tag, *h* stands for the head word, *d* stands for the dependent word, *dist* stands for the dependency distance, which has five values $\{1, 2, 3-5, 6-10, 11+\}$, and *dir* stands for the dependency direction, which has two values {left, right}.

| Tasks | | Wikitionary | | | | Parallel Data | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Delex | Proj1 | ∇ | DNN | ∇ | Proj2 | ∇ | RLAP | ∇ | X-lingual |
| En2Da | 36.5 | 41.3 | 4.8 | 42.6 | 6.1 | 42.9 | 6.4 | 43.6 | 7.1 | 38.7 |
| En2De | 46.2 | 49.2 | 3.0 | 49.5 | 3.3 | 49.7 | 3.5 | 50.5 | 4.3 | 50.7 |
| En2El | 61.5 | 62.4 | 0.9 | 63.0 | 1.5 | 63.5 | 2.0 | 64.3 | 2.8 | 63.0 |
| En2Es | 52.1 | 54.5 | 2.4 | 55.7 | 3.6 | 56.2 | 4.1 | 56.3 | 4.2 | 62.9 |
| En2It | 56.4 | 57.7 | 1.3 | 59.1 | 2.7 | 59.2 | 2.8 | 60.4 | 4.0 | 68.8 |
| En2Nl | 62.0 | 64.4 | 2.4 | 65.1 | 3.1 | 64.9 | 2.9 | 66.1 | 4.1 | 54.3 |
| En2Pt | 68.7 | 71.5 | 2.8 | 72.4 | 3.7 | 71.9 | 3.2 | 72.8 | 4.1 | 71.0 |
| En2Sv | 57.8 | 61.0 | 3.2 | 61.9 | 4.1 | 62.9 | 5.1 | 63.7 | 5.9 | 56.9 |
| Average | 55.2 | 57.8 | 2.6 | 58.7 | 3.5 | 58.9 | 3.8 | 59.7 | 4.6 | 58.3 |

Table 3: Comparison results in terms of unlabeled attachment score (UAS) for the eight cross-lingual dependency parsing tasks (English is used as the source language). The evaluation results are on *all the test sentences*. The *Delex* method uses no auxiliary resource, *Proj1* and *DNN* use Wikitionary as auxiliary resource, *Proj2, RLAP*, and *X-lingual* use parallel sentences as auxiliary resources. ∇ denotes the improvements of each method over the baseline *Delex* method. The bottom row contains the average results over the eight tasks.

sentences by using the Berkeley alignment tool (Liang et al., 2006). With the word alignments, we then projected the predicted dependency relations from the source language sentences of the parallel data to the target language side, which produces a set of dependency arc instances in the target language. Finally, we constructed the partially observed word-feature matrix from these labeled data and conducted matrix completion to recover the whole matrix. For matrix completion, we used the first task *En2Da* to perform parameter selection based on the test performance. We selected $\gamma$ from $\{0.1, 1, 10\}$ and selected $\alpha$ from $\{10^3, 10^4, 10^5\}$. The selected values $\gamma = 1$ and $\alpha = 10^{-4}$ were then used for all the experiments.

## 4.3 Experimental Results

### 4.3.1 Test Results on All the Test Sentences

We first compared the proposed representation learning with annotation projection method, *RLAP*, to the following methods in our experiments: *Delex, Proj1, Proj2, DNN* and *X-lingual*. The *Delex* method is a baseline method, which replaces the language-specific word sequence with the universal POS tag sequence and then trains a delexicalized dependency parser. We listed the feature templates used in this baseline delexicalized dependency parser in Table 2. The *Proj1* and *Proj2* methods are from (Durrett et al., 2012). Durrett et al. (2012) proposed to use bilingual lexicon to learn cross-lingual features and provided two ways to construct the bilingual lexicon, one is based on Wikitionary and the other is based on unlabeled parallel sentences with observed word-level alignments. We used these two ways separately to construct the bilingual lexicon between the languages for learning cross-lingual features, which are then used as augmenting features for training delexicalized dependency parsers. We denote the Wikitionary-based method as *Proj1* and the parallel-sentence-based method as *Proj2*. The *DNN* method, developed in (Xiao and Guo,

| Tasks | Delex | Proj2 | ∇ | RLAP | ∇ | USR | PGI | PR | MLC |
|---|---|---|---|---|---|---|---|---|---|
| En2Da | 46.7 | 54.6 | 7.9 | 55.7 | 9.0 | 51.9 | 41.6 | 44.0 | - |
| En2De | 62.0 | 63.0 | 1.0 | 64.0 | 2.0 | - | - | 39.6 | 62.8 |
| En2El | 60.9 | 61.9 | 1.0 | 63.2 | 2.3 | - | - | - | 61.4 |
| En2Es | 55.2 | 58.3 | 3.1 | 59.6 | 4.4 | 67.2 | 58.4 | 62.4 | 57.3 |
| En2It | 55.5 | 56.9 | 1.4 | 58.3 | 2.8 | - | - | - | 56.2 |
| En2Nl | 60.3 | 62.5 | 2.2 | 63.7 | 3.4 | - | 45.1 | 37.9 | 62.0 |
| En2Pt | 80.2 | 84.5 | 4.3 | 85.7 | 5.5 | 71.5 | 63.0 | 47.8 | 83.8 |
| En2Sv | 73.4 | 76.0 | 2.6 | 76.4 | 3.0 | 63.3 | 58.3 | 42.2 | 74.9 |
| Average | 61.8 | 64.7 | 2.9 | 65.8 | 4.1 | - | - | - | - |

Table 4: Comparison results on the short test sentences with length of 10 or less in terms of unlabeled attachment score (UAS). ∇ denotes the improvements of each method over the baseline *Delex* method.

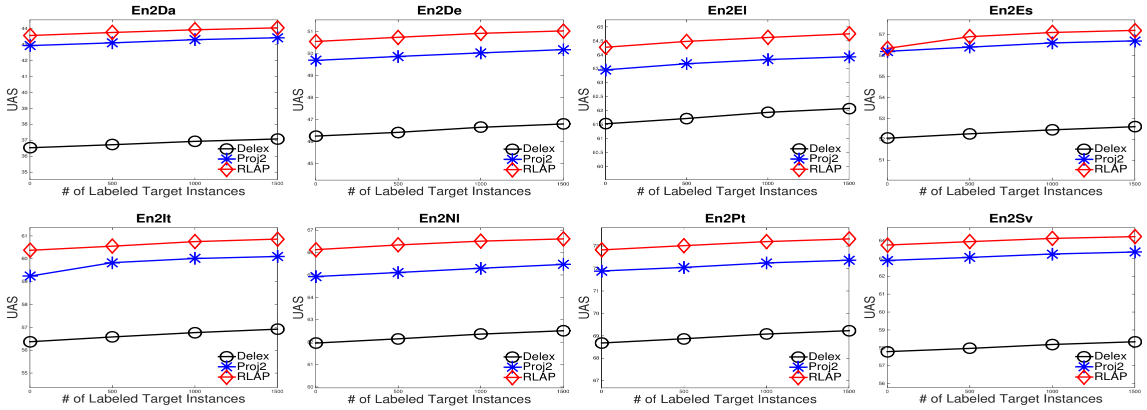

Figure 5: Unlabeled attachment score (UAS) on the whole test sentences in the target language by varying the number of labeled training sentences in the target language.

2014), uses Wikitionary to construct bilingual word pairs and then uses a deep neural network to learn interlingual word embeddings as augmenting features for training delexicalized dependency parsers. The *X-lingual* method uses unlabeled parallel sentences to induce cross-lingual word clusters as augmenting features for delexicalized dependency parser (Täckström et al., 2012). For *X-lingual*, we cited its results reported in its original paper. For other methods, we used the MSTParser (McDonald et al., 2005) as the underlying dependency parsing tool. To train the MSTParser, we set the number of maximum iterations for the perceptron training as 10 and set the number of best-k dependency tree candidates as 1.

We evaluated the empirical performance of each comparison method on all the test sentences. The comparison results on the eight cross-lingual dependency parsing tasks in terms of unlabeled attachment score (UAS) are reported in Table 3. We can see that the baseline method, *Delex*, performs poorly across the eight tasks. This is not surprising since the sequence of universal POS tags are not discriminative enough for the dependency parsing task. Note even for two sentences with the exact same sequence of POS tags, they may have different dependency trees. By using auxiliary bilingual word pairs via Wikitionary, the two cross-lingual representation learning methods, *Proj1* and *DNN*, outperform *Delex* across all the eight tasks. Between these two methods, *DNN* consistently outperforms *Proj1*, which suggests the interlingual word embeddings induced by deep neural networks are very effective. By using unlabeled parallel sentences as an auxiliary resource, the two methods, *Proj2* and *RLAP*, consistently outperform the baseline *Delex* method, while *X-lingual* outperforms *Delex* on six tasks. Moreover, *Proj2* outperforms its variant *Proj1* across all the eight tasks and achieves comparable performance with the deep neural network based method *DNN*. This suggests that unlabeled parallel sentences form a stronger auxiliary resource than the free Wiktionary. Our proposed approach, *RLAP*, which has the capacity of exploiting the unlabeled parallel sentences, consistently outperforms the four comparison methods, *Delex, Proj1, DNN* and *Proj2*, across all the eight tasks. It also outperforms the *X-lingual* method on five tasks. The average UAS over all the eight tasks for the *RLAP* method is 1.4 higher than the *X-lingual* method. All these results demonstrated the effectiveness of the proposed representation learning method for cross-lingual dependency parsing.

#### 4.3.2 Test Results on Short Test Sentences

We also conducted empirical evaluations on short test sentences (with length of 10 or less). We compared *Delex, Proj2* and *RLAP* with four other methods, *USR, PGI, PR* and *MLC*. The *USR* method is a cross-lingual direct transfer method which uses universal dependency rules to construct linguistic constraints (Naseem et al., 2010). The *PGI* method is a phylogenetic grammar induction model (Berg-Kirkpatrick and Klein, 2010). The *PR* method is a posterior regularization approach (Gillenwater et al., 2010). The *MLC* method is the multilingual linguistic constraints-based method which uses typological features for cross-lingual dependency parsing (Naseem et al., 2012). Here we used this method in our setting with only one source domain. Moreover, since we do not have typological features for Danish, we did not conduct experiment on the first task with *MLC*. For the methods of *USR, PGI* and *PR*, we cited their results reported in their original papers. All the cited results are also produced on the short sentences of the CoNLL-X shard task dataset. We cited them as references on measuring the progress of cross-lingual dependency parsing on each given target language.

The comparison results are reported in Table 4. We can see that the results on the short test sentences are in general better than on the whole test set (in Table 3) for the same method across most tasks. This suggests that it is easier to infer the dependency tree for a short sentence than for a long sentence. Nevertheless, *Proj2* consistently outperforms *Delex* and *RLAP* consistently outperforms *Proj2* across all the tasks. Moreover, *RLAP* achieves the highest test scores in seven out of the eight cross-lingual tasks among all the comparison systems. This again demonstrated the efficacy of the proposed approach for cross-lingual dependency parsing.

### 4.4 Impact of Labeled Training Data in Target Language

We have also conducted experiments for the learning scenarios where a small set of labeled training sentences from the target language is available. Specifically, we conducted experiments with a few different numbers of additional labeled training sentences from the target language, $\{500, 1000, 1500\}$, using three methods, *RLAP, Delex* and *Proj2*. The comparison results on all the test sentences are reported in Figure 5. We can see that the performance of all three methods increases very slow but in a similar trend with more additional labeled training instances from the target language. However, both *Proj2* and *RLAP* outperform *Delex* with large margins across all experiments. Moreover, the proposed method, *RLAP*, produces the best results across all the eight tasks. The results again verified the efficacy of the proposed method, demonstrated that filling the missing feature values with matrix completion is indeed useful.

## 5 Conclusion

In this paper, we proposed a novel representation learning method with annotation projection to address cross-lingual dependency parsing. The proposed approach exploits unlabeled parallel sentences and combines cross-lingual annotation projection and matrix completion-based interlingual feature learning together to automatically induce a set of language-independent numerical features. We used these interlingual features as augmenting features to train a delexicalized dependency parser on the labeled sentences in the source language and tested it in the target language domain. Our experimental results on eight cross-lingual dependency parsing tasks showed the proposed representation learning method outperforms a number of comparison methods.

## Acknowledgments

This research was supported in part by NSF grant IIS-1065397.

## References

T. Berg-Kirkpatrick and D. Klein. 2010. Phylogenetic grammar induction. In *Proc. of the Annual Meeting of the Association for Comput. Linguistics (ACL)*.

S. Buchholz and E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proc. of the Conference on Computational Natural Language Learning (CoNLL)*.

R. Cabral, F. Torre, J. Costeira, and A. Bernardino. 2011. Matrix completion for multi-label image classification. In *Advances in Neural Information Processing Systems (NIPS)*.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *J. of Machine Learning Research (JMLR)*, 12:2493–2537.

G. Durrett, A. Pauls, and D. Klein. 2012. Syntactic transfer using a bilingual lexicon. In *Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

K. Ganchev, J. Gillenwater, and B. Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proc. of the Joint Conference of the Annual Meeting of the ACL and the International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*.

J. Gillenwater, K. Ganchev, J. Graça, F. Pereira, and B. Taskar. 2010. Sparsity in dependency grammar induction. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11:11–311.

P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of the Machine Translation Summit*.

P. Liang, B. Taskar, and D. Klein. 2006. Alignment by agreement. In *Proc. of the Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*.

K. Liu, Y. Lü, W. Jiang, and Q. Liu. 2013. Bilingually-guided monolingual dependency parsing grammar induction. In *Proc. of the Conference on Annual Meeting of the Association for Computational Linguistics (ACL)*.

R. McDonald, K. Crammer, and F. Pereira. 2005. Online large-margin training of dependency parsers. In *Proc. of the Annual Meeting on Association for Computational Linguistics (ACL)*.

R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Tčkström, C. Bedini, N. Castelló, and J. Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proc. of the Annual Meeting on Association for Comput. Linguistics (ACL)*.

T. Naseem, H. Chen, R. Barzilay, and M. Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

T. Naseem, R. Barzilay, and A. Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

S. Petrov, D. Das, and R. McDonald. 2012. A universal part-of-speech tagset. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*.

D. Smith and J. Eisner. 2009. Parser adaptation and projection with quasi-synchronous grammar features. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*.

A. Søgaard and J. Wulff. 2012. An empirical study of non-lexical extensions to delexicalized transfer. In *Proc. of the Conference on Computational linguistics (COLING)*.

O. Täckström, R. McDonald, and J. Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

O. Täckström, R. McDonald, and J. Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proc. of the Conf. of the North American Chapter of the Association for Comput. Linguistics: Human Language Technologies (NAACL)*.

M. Xiao and Y. Guo. 2013. A novel two-step method for cross language representation learning. In *Advances in Neural Inform. Process. Systems (NIPS)*.

M. Xiao and Y. Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

Y. Zhang, R. Reichart, R. Barzilay, and A. Globerson. 2012. Learning to map into a universal pos tagset. In *Proc. of the Joint Conf. on Empirical Methods in Natural Language Processing and Comput. Natural Language Learning (EMNLP-CoNLL)*.

H. Zhao, Y. Song, C. Kit, and G. Zhou. 2009. Cross language dependency parsing using a bilingual lexicon. In *Proc. of the Joint Conf. of ACL and AFNLP (ACL-IJCNLP)*.