Distributed Word 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

This paper proposes to learn language-independent word representations to address cross-lingual dependency parsing, which aims to predict the dependency parsing trees for sentences in the target language by training a dependency parser with labeled sentences from a source language. We first combine all sentences from both languages to induce real-valued distributed representation of words under a deep neural network architecture, which is expected to capture semantic similarities of words not only within the same language but also across different languages. We then use the induced interlingual word representation as augmenting features to train a delexicalized dependency parser on labeled sentences in the source language and apply it to the target sentences. To investigate the effectiveness of the proposed technique, extensive experiments are conducted on cross-lingual dependency parsing tasks with nine different languages. The experimental results demonstrate the superior cross-lingual generalizability of the word representation induced by the proposed approach, comparing to alternative comparison methods.

1 Introduction

With the rapid development of linguistic resources and tools in multiple languages, it is very important to develop cross-lingual natural language processing (NLP) systems. Cross-lingual dependency parsing is the task of inferring dependency trees for observed sentences in a target language where there are few or no labeled training sentences by using a dependency parser trained on a large amount of sentences with annotated dependency trees in a source language (Durrett et al., 2012; McDonald et al., 2011; Zhao et al., 2009). Cross-lingual dependency parsing is popularly studied in natural language processing area as it can greatly reduce the expensive manual annotation effort in the target language by exploiting the dependency annotations from a source language (Durrett et al., 2012; McDonald et al., 2011; Täckström et al., 2012).

One fundamental challenge of cross-lingual dependency parsing stems from the word-level representation divergence across languages. Since sentences in different languages are expressed using different vocabularies, if we train a dependency parser on the word-level features of sentences from a source language, it will fail to parse the sentences in a different target language. A variety of work in the literature has attempted to bridge the word-level representation divergence across languages. One intuitive method delexicalizes the dependency parser by replacing the language-specific word-level features with language-independent features such as universal part-of-speech tags (Petrov et al., 2012). With the universal POS tag features, this method provides a possible way to transfer dependency parsing information from the source language to the target language and has demonstrated some good empirical results (McDonald et al., 2011). However, the number of universal POS tags is small, which limits their discriminative capacity as input features for dependency parsing. A few other works hence propose to improve the delexicalized system by learning more effective cross-lingual features such as bilingual word clusters (Täckström et al., 2012) and other interlingual representations (Durrett et al., 2012).

In this paper, we propose to address cross-lingual dependency parsing by learning distributed interlingual word representations using a deep neural network architecture. We first combine all the sentences from two language domains and

build cross language word connections based on Wikitionary, which works as a free bilingual dictionary. Then by exploiting a deep learning architecture, we learn real-valued dense feature vectors for the words in the given sentences as the high-level interlingual representations, which capture semantic similarities across languages. Finally, we use the induced distributed word representation as augmenting features to train a delexicalized dependency parser on the annotated sentences in the source language and applied it on the sentences in the target language. In order to evaluate the proposed cross-lingual learning technique, we conduct extensive experiments on eight cross-lingual dependency parsing tasks with nine different languages. The experimental results demonstrate the efficacy of the proposed approach in transferring dependency parsers across languages, comparing to other methods.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the main approach of cross-lingual word representation learning with deep neural networks and cross-lingual dependency parsing with induced interlingual features. Section 4 presents the empirical study on eight cross language dependency parsing tasks. We then conclude the paper in Section 5.

2 Related Work

Previous works developed in the literature have tackled cross-lingual dependency parsing by using cross-lingual annotation projection methods, multilingual model learning methods, and cross-lingual representation learning methods.

Cross-lingual annotation projection methods use parallel sentences to project the annotations from the source language side to the target language side and then train dependency parsers on the target data with projected annotations (Hwa et al., 2005; Liu et al., 2013; Smith and Eisner, 2009; Zhao et al., 2009). For cross-lingual annotation projection methods, both the word alignment training step and the annotation projection step can introduce errors or noise. Thus much work developed in the literature has focused on designing robust projection algorithms such as graph-based projection with label propagations (Das and Petrov, 2011), improving projection performance by using auxiliary resources such as Wikipedia metadata (Kim and Lee, 2012) or WordNet (Khapra et al., 2010), or boosting projection performance by heuristically modifying or correcting the projected annotations (Hwa et al., 2005; Kim et al., 2010). Some work has also proposed to project the discrete dependency arc instances instead of treebank as the training set (Liu et al., 2013). Moreover, besides cross-lingual dependency parsing, cross-lingual annotation projection methods have also demonstrated success in various other sequence labeling tasks including POS tagging (Das and Petrov, 2011; Yarowsky and Ngai, 2001), relation extraction (Kim et al., 2012), named entity recognition (Kim et al., 2010; Kim and Lee, 2012), constituent syntax parsing (Jiang et al., 2011), and word sense disambiguation (Khapra et al., 2010).

Multilingual model learning methods train cross-lingual dependency parsers with parameter constraints obtained from parallel data (Liu et al., 2013; Ganchev et al., 2009) or linguistic knowledges (Naseem et al., 2010; Naseem et al., 2012). Among these methods, some proposed to train a joint dependency parsing system with parameters shared across the dependency parsing models in individual languages (Liu et al., 2013). Other works used posterior regularization techniques to encode the linguistic constraints in learning dependency parsing models (Ganchev et al., 2009; Naseem et al., 2010; Naseem et al., 2012). The linguistic constraints may either come from manually constructed universal dependency parsing rules (Naseem et al., 2010) or manually specified typological features (Naseem et al., 2012), or be learned from parallel sentences (Ganchev et al., 2009). Besides cross-lingual dependency parsing, multilingual model learning methods have also achieved good empirical results for other multilingual NLP tasks, including named entity recognition (Burkett et al., 2010; Che et al., 2013; Wang and Manning, 2014), syntactic parsing (Burkett et al., 2010), semantic role labeling (Zhuang and Zong, 2010; Kozhevnikov and Titov, 2012), and word sense disambiguation (Guo and Diab, 2010).

Cross-lingual representation learning methods induce language-independent features to bridge the cross-lingual difference in the original word-level representation space and build connections across different languages. They train a dependency parser in the induced representation space by exploiting labeled data from the source language and apply it in the target language (Dur-

rett et al., 2012; Täckström et al., 2012; Zhang et al., 2012). A variety of auxiliary resources have been used to induce interlingual features, including bilingual lexicon (Durrett et al., 2012), and unlabeled parallel sentences (Täckström et al., 2013). Based on different learning mechanisms (whether or not using labeled data) for inducing language-independent features, cross-lingual representation learning methods can be categorized into unsupervised representation learning (Täckström et al., 2013) and supervised representation learning (Durrett et al., 2012). The language-independent features include bilingual word clusters (Täckström et al., 2012), language-independent projection features (Durrett et al., 2012), and automatically induced language-independent POS tags (Zhang et al., 2012). Besides cross-lingual dependency parsing, in the literature cross-lingual representation learning methods have also demonstrated efficacy in different NLP applications such as cross language named entity recognition (Täckström et al., 2012) and cross language semantic role labeling (Titov and Klementiev, 2012). Our work shares similarity with these cross-lingual representation learning methods on inducing new language-independent features, but differs from them in that we learn cross-lingual word embeddings. Though multilingual word embeddings have been employed in the literature, they are developed for other NLP tasks such as cross-lingual sentiment analysis (Klementiev et al., 2012), and machine translation (Zou et al., 2013). Moreover, the method in (Klementiev et al., 2012) requires parallel sentences with observed word-level alignments, and the method in (Zou et al., 2013) first learns language-specific word embeddings in each language separately and then transforms representations from one language to another language with machine translation alignments, while we jointly learn cross-lingual word embeddings in the two languages by only exploiting a small set of bilingual word pairs.

From the perspective of applying deep networks in natural language processing systems, there are a number of works in the literature (Collobert and Weston, 2008; Collobert et al., 2011; Henderson, 2004; Socher et al., 2011; Titov and Henderson, 2010; Turian et al., 2010). Socher et al. (2011) applied recursive autoencoders to address sentence-level sentiment classification problems. Collobert and Weston (2008) and Collobert et al. (2011) employed a deep learning framework for jointly multi-task learning and empirically evaluated it with four NLP tasks, including part-of-speech tagging, chunking, named entity recognition, and semantic role labeling. Henderson (2004) proposed discriminative training methods for learning a neural network statistical parser. Titov and Henderson (2010) extended the incremental sigmoid Belief networks (Titov and Henderson, 2007) to a generative latent variable model for dependency parsing. Turian et al. (2010) employed neural networks to induce word representations for sequence labeling tasks such as named entity recognition.

3 Cross-Lingual Dependency Parsing with Word Representation Learning

In this work, we aim to tackle cross-lingual dependency parsing by learning language-independent distributed word representations with deep neural networks. We first build connections across languages using free bilingual dictionaries. Then we introduce the deep neural network framework for cross-lingual word representation learning and describe how to employ the induced dense word embeddings for cross-lingual dependency parsing.

3.1 Building Cross Language Connections

To induce cross-lingual word representations, we first need to build connections between the source and target languages. In this work, we produce such connections by finding cross-lingual word pairs using the Wikitionary[1], which works as free bilingual dictionaries between language pairs.

Specifically, we first constructed a source language dictionary with all words that appeared in the sentences from the source language domain and translate these words to the target language using the Wikitionary. Then we filtered the produced word-to-word translations by dropping the ones where either the same source language word has multiple different word translations in the target language or the same target language word corresponds to multiple different source language words. We further dropped the word pairs where the translated word in the target language does not appear in the given sentences in the target language domain. After the processing, we have a set of one-to-one bilingual word pairs to build connections between the two language domains. Finally, we built a unified bilingual vocabulary V

[1] http://en.wikitionary.org

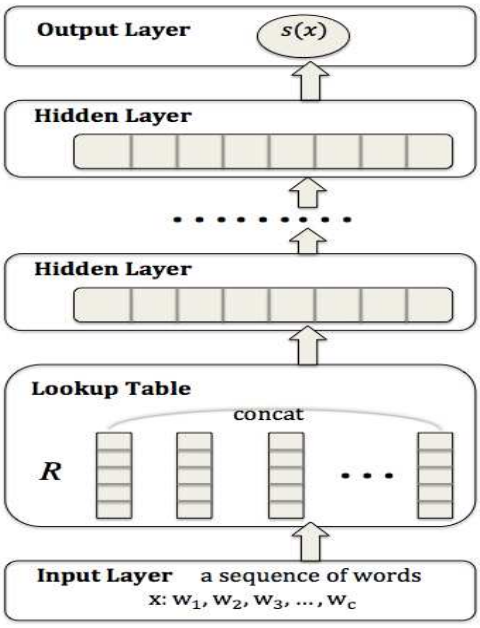

Figure 1: The architecture of the deep neural network for learning cross-lingual word representations. Each word w_i from the training sample $\mathbf{x}$ is mapped to an interlingual representation vector $R(w_i)$ through the embedding matrix R.

with words from all sentences of the two language domains. For each one-to-one bilingual word pair we constructed, we assume the two words have equivalent semantic meaning and map them to the same entry in V. Next we will learn a distributed vector representation for each entry of the bilingual vocabulary V using deep neural networks. By sharing the same representation vectors, the constructed bilingual word pairs will serve as the bridge across languages.

3.2 Interlingual Word Representation Learning with Deep Neural Networks

Given the constructed bilingual vocabulary V with v entries, we will learn a latent word embedding matrix $R \in \mathbb{R}^{k \times v}$ over the sentences in the two language domains by using a deep neural network model. This embedding matrix will map each word w in the vocabulary V into a real valued representation vector $R(w)$ with length k. For each bilingual pair of words that are mapped into the same entry of V, they will be mapped into the same vector in R as well. Following the strategy of (Collobert et al., 2011), we construct a simple two-class classification problem over the given sentences. We use the sub-sentences with fixed window size c constructed from the given sentences in the two language domains as positive samples and construct the negative samples by replacing the middle word of each positive subsentence with a random word from V. We then train a deep neural network for this two-class classification problem, while simultaneously learning the latent embedding matrix R.

The deep neural network architecture is given in Figure 1. The bottom layer of the deep architecture is the input layer, which takes a sequence of word tokens, $\mathbf{x} = w_1, w_2, \ldots, w_c$, with a fixed window size c as the input instance. Then we map each word w_i in this sequence to an embedding vector $R(w_i)$ by treating the bilingual embedding matrix R as a look-up table. The embedding vectors of the sequence of words $\mathbf{x}$ will be concatenated into a long vector $R(\mathbf{x}) \in \mathbb{R}^{ck}$ such that

$$R(\mathbf{x}) = [R(w_1); R(w_2); \ldots; R(w_c)]. \quad (1)$$

$R(\mathbf{x})$ will then be used as input for the hidden layer above it. The deep neural network has multiple hidden layers. The first hidden layer applies a nonlinear hyperbolic tangent activation function over the linear transformation of its input vector $R(\mathbf{x})$, such that

$$H_1(\mathbf{x}) = \tanh\left(W_1 \times R(\mathbf{x}) + \mathbf{b}_1\right) \quad (2)$$

where $W_1 \in \mathbb{R}^{h_1 \times ck}$ is the weight parameter matrix, $\mathbf{b}_1 \in \mathbb{R}^{h_1}$ is the bias parameter vector, $H_1(\mathbf{x}) \in \mathbb{R}^{h_1}$ is the output vector, and h_1 is the number of hidden units in the first hidden layer. Similarly, each of the other hidden layers takes the previous layer's output as its input and performs a nonlinear transformation to produce an output vector. For example, for the i-th hidden layer, we used $H_{i-1}(\mathbf{x})$ as its input and $H_i(\mathbf{x})$ as its output such that

$$H_i(\mathbf{x}) = \tanh\left(W_i \times H_{i-1}(\mathbf{x}) + \mathbf{b}_i\right) \quad (3)$$

where $W_i \in \mathbb{R}^{h_i \times h_{i-1}}$ is the weight parameter matrix and $\mathbf{b}_i$ is the bias parameter vector for the i-th hidden layer; h_i denotes the number of hidden units of the i-th hidden layer.

Given t hidden layers, the output representation of the last layer will then be used to generate a final score value for the prediction task, such that

$$s(\mathbf{x}) = \theta \times H_t(\mathbf{x}) + u \quad (4)$$

where $\theta \in \mathbb{R}^{h_t}$ is the weight parameter vector and u is the bias parameter for the output layer.

In summary, the model parameters of the deep neural network architecture include the look-up table R, the parameters $\{W_i, \mathbf{b}_i\}_{i=1}^{t}$ for the hidden layers, and the output layer parameters (θ, u).

3.3 The Training Procedure

The model parameters of the deep network architecture are learned by training a two-class classification model over the the constructed positive and negative samples. Let $D = \{\mathbf{x}_i, \hat{\mathbf{x}}_i\}_{i=1}^{N}$ denote the constructed training set, where $\mathbf{x}_i$ is a positive sample and $\hat{\mathbf{x}}_i$ is a negative sample constructed by replacing the middle word of $\mathbf{x}_i$ with a random word from V. It is desirable for the model to produce an output score $s(\mathbf{x}_i)$ that is much larger than the score $s(\hat{\mathbf{x}}_i)$ for each pair of training instances. Thus we perform training to maximize the separation margins between the pairs of scores over positive and negative samples under a hinge loss; that is we minimize the following training loss

$$J(D) = \frac{1}{N} \sum_{i=1}^{N} \max(0, 1 - s(\mathbf{x}_i) + s(\hat{\mathbf{x}}_i)) \quad (5)$$

We perform a random initialization over the look-up table and weight model parameters, and set all the bias model parameters to zeros. Then we use a stochastic gradient descent (Bottou, 1991) algorithm to perform optimization.

3.4 Cross-Lingual Dependency Parsing

The training of deep network model above will produce a word embedding matrix R for all words in the two language domains. Moreover, by having each translated bilingual pair of words sharing the same representation vector in R in the training process, the embedding matrix R is expected to capture consistent and comparable semantic meanings across languages, and provide a language-independent and distributed representation for each word in the bilingual dictionary V.

Given R, for each sentence $\mathbf{x} = w_1, w_2, \ldots, w_n$ from the two language domains, we retrieved the representation vector $R(w_i)$ for each word w_i. Moreover, we further delexicalized the sentence by replacing the sequence of language-specific words with a sequence of universal POS tags (Petrov et al., 2012). Finally we train a delexicalized dependency parser on the labeled sentences in the source language based on the universal POS tag features and the learned distributed features. and apply it to perform dependency parsing on the sentences in the target language domain.

4 Experiments

We empirically evaluated the proposed cross-lingual word representation learning for cross-lingual dependency parsing. In this section, we present the experimental setup and the results.

4.1 Dataset

We used the dataset from the CoNLL shared task (Buchholz and Marsi, 2006; Nivre et al., 2007) for cross-lingual dependency parsing. We conducted experiments with the following nine languages: English (EN), Danish (DA), German (DE), Greek (EL), Spanish (ES), Italian (IT), Dutch (NL), Portuguese (PT) and Swedish (SV). For each language, there is a separate training set and a test set. We used English, which usually has more labeled resources, as the source language, while treating the others as target languages. We thus constructed eight cross-lingual dependency parsing tasks (EN2DA, EN2DE, EN2EL, EN2ES, EN2IT, EN2NL, EN2PT, EN2SV), one for each of the eight target languages. For example, the task *EN2DA* means that we used Danish (DA) as the target language while using *English (EN)* as the source language. For each cross language dependency parsing task, we first performed representation learning and then conducted dependency parsing training and test.

In this dataset, each sentence is labeled with gold standard part-of-speech tags. To produce delexicalized cross-lingual dependency parsers, we mapped these language-specific part-of-speech tags into twelve universal POS tags (Petrov et al., 2012): ADJ (adjectives), ADP (prepositions or postpositions), ADV (adverbs), CONJ (conjunctions), DET (determiners), NOUN (nouns), NUM (numerals), PRON (pronouns), PRT (particles), PUNC (punctuation marks), VERB (verbs) and X (for others).

4.2 Representation Learning

For each language pair, we produced a set of one-to-one bilingual word pairs using Wikitionary to build cross language connections. The numbers of bilingual word pairs produced for all the eight language pairs and the numbers of words in each language are given in Table 1.

Table 1: The number of words in each language and the number of selected bilingual word pairs for each of the eight language pairs.

Language Pairs	# Source Words	# Target Words	# Bilingual Word Pairs
English vs Danish	26599	17934	1140
English vs Dutch	26599	27829	2976
English vs German	26599	69336	1905
English vs Greek	26599	13318	869
English vs Italian	26599	13523	2347
English vs Portuguese	26599	27782	2408
English vs Spanish	26599	16465	2910
English vs Swedish	26599	19072	1779

Table 2: The feature templates used for the cross-lingual dependency parsing. dir denotes the direction of the dependency relationship, which has two values $\{left, right\}$. $dist$ denotes the distance between the head word and the dependent word, which has five values {1, 2, 3-5, 6-10, 11+}.

Feature Template	Feature Description
$UPOS(w_h)$	the head word's universal POS tag
$UPOS(w_d)$	the dependent word's universal POS tag
$UPOS(w_h, w_d)$	the universal POS tag pair of the head and dependent word
$R(w_h)$	the head word's distributed representation
$R(w_d)$	the dependent word's distributed representation
$dir\&UPOS$	conjunction features related to the dependency direction
$dist\&UPOS$	conjunction features related to the dependency distance
$dir\&dist\&UPOS$	conjunction features related to the dependency direction and distance

To perform distributed cross-lingual representation learning using the proposed deep network architecture, we first constructed the two-class training dataset from all the sentences (training and test sentences) of the two language domains. This requires the creation of sub-sentences with fixed window size c from the given sentences. We used window size $c = 5$ in the experiments. For example, for a given sentence "I visited New York ." , we can produce a number of sub-sentences, including "<PAD> <S> I visited New", "<S> I visited New York", "I visited New York .", "visited New York . </S>", and "New York . </S> <PAD>", where <PAD> is special token to fill the length requirement. Negative samples are constructed by simply replace the middle word of each sub-sentence with a random word.

With the constructed training data, we then performed training over the deep neural network. We used 3 hidden layers with 100 hidden units in each layer, considering the model capacity and the training effort. The dimension k of the embedding word vectors in R is set as 200.

4.3 Cross-lingual Dependency Parsing

We used the MSTParser (McDonald et al., 2005a; McDonald et al., 2005b) as the basic dependency parsing model. MSTParser uses spanning tree algorithms to seek for the candidate dependency trees and employs an online large margin training optimization algorithm. MSTParser is widely used in the literature for dependency parsing tasks and has demonstrated good empirical results in the CoNLL shared tasks on multilingual dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007). For this dependency parsing model, there are a few parameters to be set: the number of maximum iterations for the perceptron training, and the number of best-k dependency tree candidates. We set the number of iterations to be 10 and only considered the best-1 dependency tree candidate.

For the proposed cross-lingual dependency parsing approach, we used both the delexi-

Table 3: Test performance in terms of UAS (unlabeled attachment score) on the eight cross-lingual dependency parsing tasks. Δ denotes the improvements of each method over the *Baseline* method.

Tasks	Baseline	Proj	Δ	Proposed	Δ	X-lingual
EN2DA	36.53	41.25	4.72	42.56	6.03	38.70
EN2DE	46.24	49.15	2.91	49.54	3.30	50.70
EN2EL	61.53	62.36	0.83	62.96	1.43	63.00
EN2ES	52.05	54.54	2.49	55.72	3.67	62.90
EN2IT	56.37	57.71	1.34	59.05	2.68	68.80
EN2NL	61.96	64.41	2.45	65.13	3.17	54.30
EN2PT	68.68	71.47	2.79	72.38	3.70	71.00
EN2SV	57.79	60.99	3.20	61.88	4.09	56.90
Average	55.14	57.74	2.60	58.90	3.51	58.29

calized universal POS tag based features and the language-independent word features produced from the deep learning as input features for the MSTParser. The set of universal POS tag based feature templates is given in Table 2. For each dependency relationship between a head word w_h and a dependent word w_d, a set of features can be produced from the feature templates in Table 2, which can be further augmented by $R(w_h)$ and $R(w_d)$. We compared our proposed approach (*Proposed*) with three other methods, *Baseline*, *Proj* and *X-lingual*. The *Baseline* method uses a delexicalized MSTParser based only on the universal POS tag features. The *Proj* method is developed in (Durrett et al., 2012), which uses a bilingual dictionary to learn cross-lingual features and then uses them as augmenting features to train a delexicalized MSTParser. The *X-lingual* method uses unlabeled parallel sentences to learn cross-lingual word clusters and used them as augmenting features to train a delexicalized MSTParser (Täckström et al., 2012). All parsers except *X-lingual* are trained on the labeled sentences in the source language domain and tested on the test sentences in the target language domain in the given dataset. The performance is measured using the standard unlabeled attachment score (UAS). The *X-lingual* method uses different auxiliary resources (parallel sentences), and we hence directly cited the results reported in (Täckström et al., 2012) on the same dataset.

4.4 Results and Discussions

We reported the empirical comparison results in terms of unlabeled attachment score (UAS) in Table 3. We can see that the *Baseline* method performs poorly across all the tasks. The average unlabeled attachment score for this approach across all the eight tasks is very low (about 55.14), which suggests that the twelve universal POS tags are far from enough to produce a good cross-lingual dependency parser. Considering the small number of universal POS tags, its limited discriminative capacity as input features for dependency parsing is understandable. To further verify this, we calculated the percentage of sentences in the test data which share the same sequence of universal POS tags with a training sentence in the source language but have different dependency parsing structures. The values for the eight tasks are presented in Table 4. The non-trivial values reported verified the universal POS tags' drawback on lacking discriminative capacity.

Table 4: Statistic differences. For each task, we report the percentage of sentences in the test data from the target language which share the same sequence of universal POS tags with some sentences in the source language but with different dependency trees.

Target Language	Sentence Difference
Danish	0.31%
Dutch	1.81%
German	1.40%
Greek	1.20%
Italian	2.40%
Portuguese	1.04%
Spanish	0.97%
Swedish	2.31%

By *relexicalizing* the delexicalized MSTParser

via augmenting the POS tag sequences with learned interlingual features, both the *Proj* method and the proposed method overcome the drawback of using solely universal POS tags and produce significant improvements over the *Baseline* method across all the tasks. Moreover, the proposed method consistently outperforms both *Baseline* and *Proj* for all the eight tasks. By exploiting only free bilingual dictionaries, the proposed method achieves similar average performance to the *X-lingual* method which requires additional parallel sentences. All these results demonstrated the efficacy of our word representation learning method for cross-lingual dependency parsing.

4.5 Impact of Labeled Training Data in Target Language

In the experiments above, all the labeled sentences for dependency parsing training are from the source language. We wonder how much benefit we can get if there are a small number of labeled sentences in the target language as well. To answer this question, we conducted experiments by using a small number (ℓ_t) of labeled sentences in the target language domain together with the labeled sentences in the source language domain to train cross-lingual dependency parsers. Again the performance of the parsers are evaluated on the test sentences in the target language. We tested a few different ℓ_t values with $\ell_t \in \{500, 1000, 1500\}$. We reported the unlabeled attachment score for all the eight cross-lingual dependency parsing tasks in Figure 2. We can see that the *Baseline* method still performs poorly across the range of different setting for all the eight tasks. The *Proj* method and the proposed method again consistently outperform the baseline method across all the tasks, while the proposed method achieves the best results across all the eight tasks.

4.6 Impact of the Number of Bilingual Word Pairs

For the eight language pairs, we have reported the numbers of words in each language domain and the numbers of selected bilingual word pairs in Table 1. Next we investigated how the number of word pairs affects the performance of the proposed cross-lingual dependency parsing. With the selected full set of bilingual word pairs in Table 1, we random selected $m\%$ of them with $m \in \{50, 75, 100\}$ to conduct experiments. Note when $m = 50$, we only used 435 word pairs for the EN2EL (English vs. Greek) task, which is 1.6% of the number of source words and 3.3% of the number of target words. The results are reported in Figure 3. We can see that by reducing the number of bilingual word pairs, the performance of the proposed cross-lingual dependency parsing method degrades on all tasks. This is reasonable since the word pairs serve as the pivots for learning cross-lingual word embeddings. Nevertheless, by preserving 75% of the selected word pairs, the proposed approach can still outperform the *Proj* method across all the tasks. Even with only 50% of the word pairs, our method still outperforms the *Proj* method on most tasks. These results suggest that the proposed cross-lingual word embedding method only requires a reasonable amount of bilingual word pairs to effectively transfer a dependency parser from the source language to the target language.

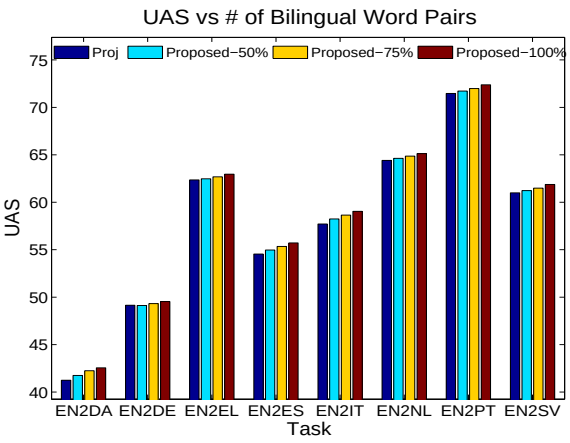

Figure 3: Test performance in terms of UAS (unlabeled attachment score) in the target language with different numbers of bilingual word pairs.

5 Conclusion

In this paper, we proposed to automatically learn language-independent features within a deep neural network architecture to address cross-lingual dependency parsing problems. We first constructed a set of bilingual word pairs with Wiktionary, which serve as the pivots in the bilingual vocabulary for building connections across languages. We then conducted distributed word representation learning by training a constructed auxiliary classifier using deep neural networks, which induced a real-valued embedding vector for each word of the bilingual vocabulary to capture con-

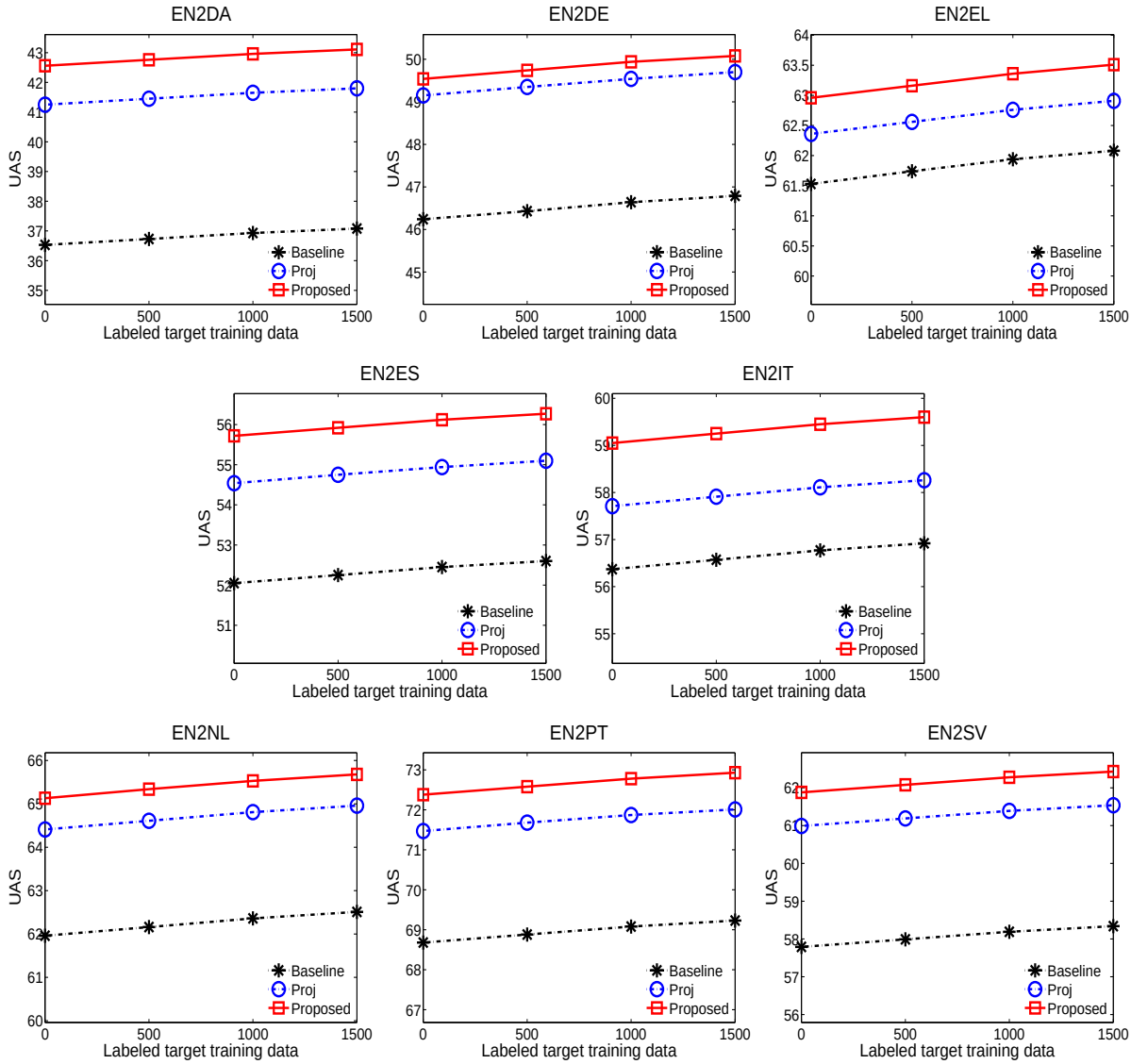

Figure 2: Unlabeled attachment score (UAS) on the test sentences in the target language by using different number of additional labeled training sentences in the target language.

sistent semantic similarities for words in the two language domains. The distributed word embedding vectors were then used to augment the universal POS tags to train cross-lingual dependency parsers. We empirically evaluated the proposed method on eight cross-lingual dependency parsing tasks between eight language pairs. The experimental results demonstrated the effectiveness of the proposed method, comparing to other cross-lingual dependency parsing methods.

References

L. Bottou. 1991. Stochastic gradient learning in neural networks. In *Proceedings of Neuro-Nîmes*.

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

D. Burkett, S. Petrov, J. Blitzer, and D. Klein. 2010. Learning better monolingual models with unannotated bilingual text. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

W. Che, M. Wang, C. Manning, and T. Liu. 2013. Named entity recognition with bilingual constraints. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: Deep neural

networks with multitask learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

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

D. Das and S. Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*.

G. Durrett, A. Pauls, and D. Klein. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings 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 *Proceedings 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)*.

W. Guo and M. Diab. 2010. Combining orthogonal monolingual and multilingual sources of evidence for all words wsd. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

J. Henderson. 2004. Discriminative training of a neural network statistical parser. In *Proceedings 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.

W. Jiang, Q. Liu, and Y. Lü. 2011. Relaxed cross-lingual projection of constituent syntax. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

M. Khapra, S. Sohoney, A. Kulkarni, and P. Bhattacharyya. 2010. Value for money: Balancing annotation effort, lexicon building and accuracy for multilingual wsd. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

S. Kim and G. Lee. 2012. A graph-based cross-lingual projection approach for weakly supervised relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

S. Kim, M. Jeong, J. Lee, and G. Lee. 2010. A cross-lingual annotation projection approach for relation detection. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

S. Kim, K. Toutanova, and H. Yu. 2012. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

A. Klementiev, I. Titov, and B. Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

M. Kozhevnikov and I. Titov. 2012. Cross-lingual bootstrapping for semantic role labeling. In *Proceedings of the NIPS Workshop on Crosslingual Technologies (XLITE)*.

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

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

R. McDonald, F. Pereira, K. Ribarov, and J. Hajič. 2005b. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*.

R. McDonald, S. Petrov, and K. Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

R. Socher, J. Pennington, E. Huang, A. Ng, and C. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

I. Titov and J. Henderson. 2007. Constituent parsing with incremental sigmoid belief networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

I. Titov and J. Henderson. 2010. A latent variable model for generative dependency parsing. In *Proceedings of the International Conference on Parsing Technology (IWPT)*.

I. Titov and A. Klementiev. 2012. Crosslingual induction of semantic roles. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

J. Turian, L. Ratinov, and Y. Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

M. Wang and C. Manning. 2014. Cross-lingual pseudo-projected expectation regularization for weakly supervised learning. *Transactions of the Association for Computational Linguistics (TACL)*, 2:55–66.

D. Yarowsky and G. Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies (NAACL)*.

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

H. Zhao, Y. Song, C. Kit, and G. Zhou. 2009. Cross language dependency parsing using a bilingual lexicon. In *Proceedings 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)*.

T. Zhuang and C. Zong. 2010. Joint inference for bilingual semantic role labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

W. Zou, R. Socher, D. Cer, and C. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.