Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

# A Representation Learning Framework for Multi-Source Transfer Parsing

**Jiang Guo**[1*], **Wanxiang Che**[1†], **David Yarowsky**[2], **Haifeng Wang**[3], **Ting Liu**[1]
[1]Center for Social Computing and Information Retrieval Harbin Institute of Technology, Harbin, China
[2]Center for Language and Speech Processing Johns Hopkins University, Baltimore, USA
[3]Baidu Inc., Beijing, China
{jguo, car, tliu}@ir.hit.edu.cn
yarowsky@jhu.edu, wanghaifeng@baidu.com

## Abstract

Cross-lingual model transfer has been a promising approach for inducing dependency parsers for low-resource languages where annotated treebanks are not available. The major obstacles for the model transfer approach are two-fold: 1. Lexical features are not directly transferable across languages; 2. Target language-specific syntactic structures are difficult to be recovered. To address these two challenges, we present a novel representation learning framework for multi-source transfer parsing. Our framework allows multi-source transfer parsing using full lexical features straightforwardly. By evaluating on the Google universal dependency treebanks (v2.0), our best models yield an absolute improvement of 6.53% in averaged labeled attachment score, as compared with delexicalized multi-source transfer models. We also significantly outperform the state-of-the-art transfer system proposed most recently.

## Introduction

The goal of dependency parsing is to induce tree structures for natural language sentences following the dependency grammar. Dependency parsing can be highly beneficial for various natural language processing (NLP) tasks, such as question answering, machine translation, knowledge mining/representation. Most of the previous work on dependency parsing focused on supervised learning with human-annotated treebanks, which are limited to very few resource-rich languages such as English and Chinese. Since it is labor intensive and time-consuming to manually build treebanks for all languages, recent years have seen a great deal of interest in cross-lingual learning methods which aim at inducing dependency parsers for low-resource languages while using only annotated training data from resource-rich languages.

Several approaches have been proposed for cross-lingual dependency parsing, mainly including annotation projection methods (Hwa et al. 2005; Tiedemann 2014) and model transfer methods (McDonald, Petrov, and Hall 2011). Another line of research on multilingual dependency parsing is unsupervised grammar induction (Klein and Manning 2004). However, in terms of accuracy, the annotation projection approach and the model transfer approach are far more promising than the unsupervised ones. In this study, we investigate the cross-lingual model transfer approach.

The pioneering work on model transfer of dependency parsing is the delexicalized models of McDonald, Petrov, and Hall (2011), which ignore all lexical features that are not directly transferable across languages. To address the deficiency regarding the lexical features, Täckström, McDonald, and Uszkoreit (2012) proposed cross-lingual word clusters, which can be viewed as partial lexical feature representations. More recently, Guo et al. (2015) proposed to fill the lexical feature gap by learning bilingual word embeddings. They combined bilingual word embeddings and word clusters, showing significant improvement against the delexicalized systems.

However, the major problem of the aforementioned representation learning approaches is that they only support transfer between two languages. Therefore, the resulting models lack the ability of recovering some specific syntactic structures of the target language which rarely (or never) appear in the source language. Take the word ordering as an example, in Spanish and French, *adjectives* often appear after *nouns*, yielding right-directed *amod* (adjective modifier) dependency arcs, whereas in English (source language) most of the *amod* arcs are left-directed. Another typical example is German, in which verbs appear mostly in V2 position, yielding left-directed *dobj* (direct object) arcs, which can hardly be recovered using models trained in English.

So what if we include a language which is more syntactically similar to the target language as one of our source languages? This intuition results in multi-source transfer parsing, which can significantly improve the overall quality of the resulting parsers (McDonald, Petrov, and Hall 2011). To this end, we propose a novel representation learning framework for multi-source transfer parsing that integrates both ideas to improve cross-lingual model transfer.

The main challenge that arises is to learn unified word representations over multiple languages. We present two algorithms for learning unified word embeddings, namely **multilingual skip-gram** (MULTI-SG) and **multilingual robust projection** (MULTI-PROJ), which are respectively extensions of the monolingual skip-gram model (Mikolov et al. 2013) and the robust projection approach proposed by Guo

[*] This work was done while the author was visiting JHU.
[†] Email corresponding.

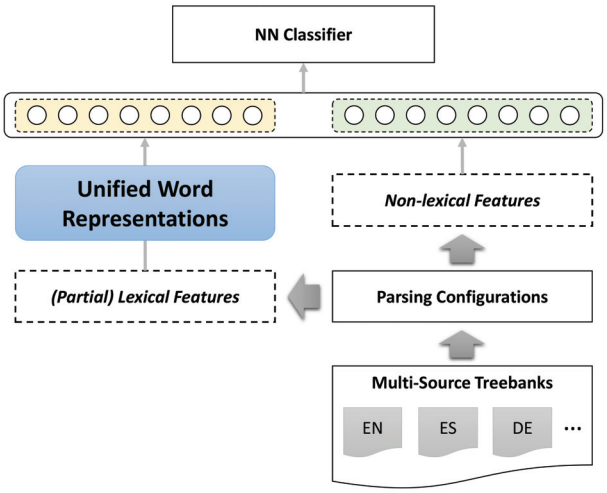


Figure 1: The architecture for multi-source transfer parsing.

et al. (2015). To make effective use of the distributed feature representations, we employ the non-linear neural network architecture for transition-based dependency parsing proposed by Chen and Manning (2014) as the basis of our transfer parsing system. As illustrated in Figure 1, treebanks from multiple source languages are concatenated as training data. The induced unified word representations are used for transforming the multilingual lexical features to a shared embedding space as input to a neural network classifier.

We consider all Indo-European languages presented in the universal dependency treebanks (v2.0) (McDonald et al. 2013), to evaluate our approach.[1] Experiment results are promising: our best models improve upon the multi-source delexicalized transfer by 6.53% of averaged labeled attachment score (LAS). We also outperform the state-of-the-art transfer system proposed most recently (Zhang and Barzilay 2015). We further show that our framework can readily incorporate minimal supervision from the target languages (50 annotated sentences) to boost the performance. Our original major contributions in this paper include:

- We propose two novel and effective approaches for learning unified word embeddings across multiple languages.
- We propose a representation learning framework for multi-source cross-lingual transfer parsing, and demonstrate its effectiveness on a benchmark dataset.

## Background

### Dependency Parsing

As a long-standing central problem in NLP, dependency parsing has attracted a great deal of interest during the last two decades. Formally, given an input sentence $\mathbf{x} = w_0 w_1 ... w_n$, the goal of dependency parsing is to build a dependency tree (Figure 2), denoted by $\mathbf{d} = \{(h, m, l) : 0 \leq h \leq n; 0 < m \leq n, l \in \mathcal{L}\}$. $(h, m, l)$ indicates a directed dependency arc from the head word $w_h$ to the modifier $w_m$ with a dependency relation $l$, and $\mathcal{L}$ is the relation set.

[1]English (EN), German (DE), Spanish (ES), French (FR), Italian (IT), Portuguese (PT) and Swedish (SV).

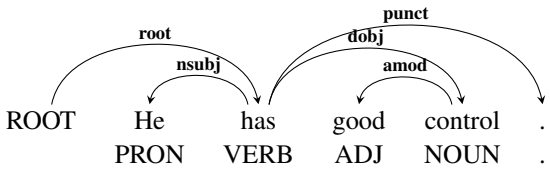


Figure 2: An example labeled dependency tree.

Various supervised models for monolingual parsing have been proposed, primarily including graph-based models and transition-based models (McDonald and Nivre 2007). From the perspective of feature representations, we can also describe the parsing models as the traditional discrete representation-based models and the distributed representation-based models that utilize non-linear neural networks for structure prediction (Chen and Manning 2014; Dyer et al. 2015; Weiss et al. 2015; Zhou et al. 2015). In this study, we adopt the distributed representation-based model which is favorable for our framework.

### Multilingual Transfer Parsing

The lack of annotated parsing resources for the vast majority of world languages has given rise to a line of research on cross-lingual transfer parsing. There are two typical categories of approaches for transfer parsing along two orthogonal directions, namely annotation projection approaches and model transfer approaches.

The basic principle of the annotation projection approaches is to project linguistic structures from a resource-rich language to a resource-poor language via bilingual word-aligned sentence pairs. As a result, a noised automatically annotated treebank will be constructed for the target language, which are then used for supervised training (Hwa et al. 2005; Tiedemann 2014).

Model transfer, however, doesn't assume/require access to bilingual parallel data. Models are trained from single or multiple source language treebanks and then applied directly to the target language of interest. Along this line, multi-source transfer has been shown to be highly beneficial (McDonald, Petrov, and Hall 2011; Zhang and Barzilay 2015). On the other hand, feature representations turn out to be critical for model transfer on more fine-grained tasks (labeled parsing) (Guo et al. 2015; Duong et al. 2015). The idea of combining these two contributions kindles this study.

## Our Framework

This section describes the two primary components of our framework.

### Learning Unified Word Representations

Most previous approaches of learning cross-lingual word representations focused solely on bilingual scenario (Klementiev, Titov, and Bhattarai 2012; Zou et al. 2013; Xiao and Guo 2014; Chandar et al. 2014; Hermann and Blunsom

2014; Faruqui and Dyer 2014; Gouws, Bengio, and Corrado 2014; Lu et al. 2015; Luong, Pham, and Manning 2015; Guo et al. 2015). Our framework, however, should be able to handle multiple languages to support multi-source transfer. Therefore, the first component of this framework is learning unified multilingual word representations.

Quite a few recent work proposed to learn cross-lingual word embeddings without word alignment information, such as the bilingual autoencoder approach (Chandar et al. 2014) and the compositional vector models (Hermann and Blunsom 2014). Those models are mostly evaluated on cross-lingual document classification tasks, where topic-related similarities are favorable. On the contrary, we emphasize the importance of alignment information here, since word-to-word translational equivalence is important for syntactic tasks, in which predictions are targeted on word units.

We present two algorithms of learning unified multilingual word representations, i.e. word embeddings. First, we consider a natural extension of the skip-gram model (Mikolov et al. 2013) as implemented in the well-known *word2vec* toolkit to multilingual scenario. Then, we present a multilingual robust projection approach.

**Model 1: Multilingual Skip-gram** Among various neural network language models, the skip-gram model has attracted a great deal of interest in recent years, due to its simplicity of implementation, efficiency of training and efficacy on many practical tasks. Here we briefly review the basic skip-gram model of learning monolingual word embeddings. The model takes the current word $w$ as input, and predicts the context words surrounding it. Denoting the word embedding of $w$ as $\mathbf{v}_w$ and context embedding of $c$ as $\mathbf{v}'_c$, the probability distribution of $c$ given $w$ follows a $softmax$ function:

$$p(c|w;\theta) = \frac{exp(\mathbf{v}'^{\top}_c \mathbf{v}_w)}{\sum_{c' \in V} exp(\mathbf{v}'^{\top}_{c'} \mathbf{v}_w)} \quad (1)$$

where $V$ is the vocabulary, and the parameters $\theta$ include the word embedding matrix and the context embedding matrix. The model can be trained by maximizing the log-likelihood over the entire training data $D$ which is the set of all word-context pairs:

$$J(\theta) = \sum_{(w,c) \in D} \log p(c|w;\theta) \quad (2)$$

We present a natural extension of this model to learn multilingual word embeddings. Recall the distributional hypothesis (Firth 1957):*"You shall know a word by the company it keeps"*, which indicates that if two words have the same/similar meaning, they are expected to share similar context words. We suggest that this hypothesis hold for multilingual words as well. Therefore, it is intuitive to predict context words *cross-lingually* based on word alignments.

We assume the access to bilingual parallel data between English and each of the other languages. First, we conduct unsupervised word alignment for each bilingual parallel data, which bridges the words across languages. Take EN/FR/ES as a case study, as shown in Figure 3, ⟨**accepter**, **accept**⟩ and ⟨**accept**, **acceptor**⟩ are aligned

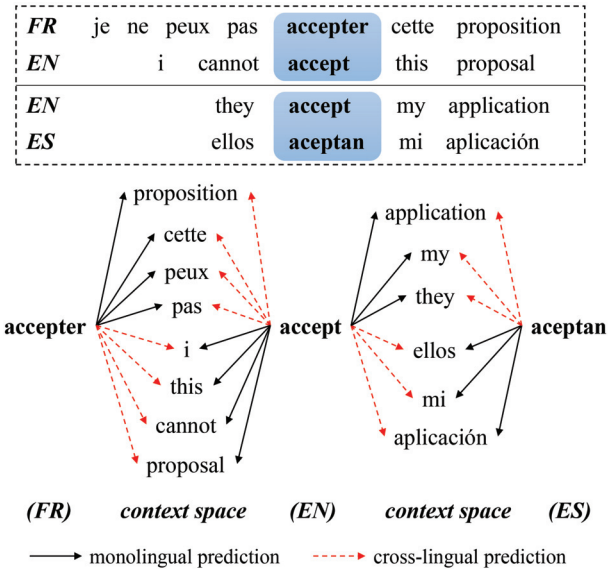


Figure 3: An example of the multilingual skip-gram model using window size of 5, taking EN/FR/ES as a case study.

word pairs in EN/FR and EN/ES parallel sentence respectively. In the multilingual skip-gram model, we include both monolingual and cross-lingual contexts for prediction. Hence the training data $D$ will be consisting of both monolingual and cross-lingual word-context pairs: $D_{EN \leftrightarrow EN}, D_{FR \leftrightarrow FR}, D_{ES \leftrightarrow ES}, D_{EN \leftrightarrow FR}, D_{EN \leftrightarrow ES}$. The English contexts will be bridges drawing the embeddings of **accepter** and **acceptor** to be close.

Denoting the set of languages as $\mathcal{L}$, we examine the following joint objective:

$$\begin{aligned} J &= \alpha \sum_{l \in \mathcal{L}} J_{mono_l} + \beta \sum_{l \in \mathcal{L} \setminus \{EN\}} J_{bi_{l,EN}} \\ J_{mono_l} &= \sum_{(w,c) \in D_{l \leftrightarrow l}} \log p(c|w;\theta) \\ J_{bi_{l,EN}} &= \sum_{(w,c) \in D_{EN \leftrightarrow l}} \log p(c|w;\theta) \end{aligned} \quad (3)$$

$J_{mono}$ and $J_{bi}$ utilize the same formulation as the basic skip-gram model, except that $J_{bi}$ use the cross-lingual contexts. $\alpha$ and $\beta$ can be tuned in practice. In this paper, we simply set them equally to 1. The model is trained using the negative sampling algorithm (Mikolov et al. 2013).

Note that $J_{mono_l}$ can be derived from both bilingual parallel corpus and additional monolingual data, which makes this approach flexible to utilizing much richer resources.

**Model 2: Multilingual Robust Projection** Next, we consider the robust projection algorithm, which has proven effective in bilingual transfer of dependency parsing models (Guo et al. 2015). We present here an extension of the robust projection algorithm to multilingual scenario.

As demonstrated in Figure 4, the basic robust projection is conducted independently for each language from English,[2]

[2]Throughout this work, we choose English as the seeding lan-

which can be formalized as a pipeline of two stages, namely bilingual propagation and monolingual propagation.

**Bilingual propagation**. Take Spanish as an example. We first collect a bilingual alignment matrix $A_{ES|EN} \in \mathbb{R}^{|V_{ES}|\times|V_{EN}|}$ from bitexts where $V_{ES}$ and $V_{EN}$ are vocabularies for EN and ES respectively. Each element of $A_{ES|EN}$ is a normalized count of alignments between corresponding words in each vocabulary:

$$A_{ES|EN}(i,j) = \frac{\#(V_{ES}^{(i)} \leftrightarrow V_{EN}^{(j)})}{\sum_k \#(V_{ES}^{(i)} \leftrightarrow V_{EN}^{(k)})} \quad (4)$$

Given a pre-trained English (the seeding language) word embedding matrix $E_{EN}$, the resulting word embedding matrix for ES can be simply computed as:

$$E_{ES}^{in} = A_{ES|EN} \cdot E_{EN} \quad (5)$$

Therefore, the embedding of each word in ES is the weighted average of the embeddings of its translation words in our bilingual parallel corpus.

Through bilingual propagation, we obtain word embeddings for each word appears within our alignment dictionary. In order to improve the word coverage, we further applied a monolingual propagation procedure to induce word embeddings for out-of-vocabulary (OOV) words.

**Monolingual propagation**. We can make effective use of various similarity measures to propagate information monolingually. In Guo et al. (2015), they utilized *edit distance* similarity to build connections between out-of-vocabulary words ($E_{ES}^{oov}$) and in-vocabulary words ($E_{ES}^{in}$). For Indo-European languages, this does make sense, thus we follow their work and construct an *edit distance* matrix $M_{ES} \in \mathbb{R}^{|V_{ES}^{oov}|\times|V_{ES}^{in}|}$ where:

$$M_{ES}(i,j) = \begin{cases} 1, & if\ editdist(V_{ES}^{oov(i)}, V_{ES}^{in(j)}) \leq \tau \\ 0, & if\ editdist(V_{ES}^{oov(i)}, V_{ES}^{in(j)}) > \tau \end{cases} \quad (6)$$

$\tau$ is set to 1. After normalizing $M$ by row, the OOV word embeddings can be simply computed as:

$$E_{ES}^{oov} = M_{ES} \cdot E_{ES}^{in} \quad (7)$$

Note that the multilingual robust projection approach can also be flexibly extended to two distant languages via bridge languages.

**Multilingual Word Clustering** Previous work (Guo et al. 2015) has demonstrated the efficacy of cross-lingual word clustering for bilingual transfer parsing. Here, we extend their approach to multilingual scenario by simply applying the PROJECTED cluster strategy (Täckström, McDonald, and Uszkoreit 2012) to each language pair independently taking English as the seeding language.

## Distributed representations-based Dependency Parsing

To make effective use of the unified distributed word representations, we employ the non-linear neural network architecture proposed by Chen and Manning (2014) as the basis of our multi-source transfer parsing system (Figure 1).

guage, and do not consider the interactions between other languages due to the unavailability of corresponding bitexts.

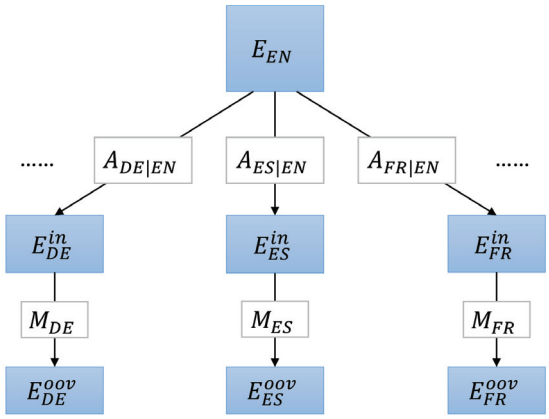


Figure 4: Multilingual robust projection.

We also adopt all improvements upon this model proposed by Guo et al. (2015), i.e. the non-local features (*distance*, *valency*) (Zhang and Nivre 2011), and the cluster features.

Revisiting the architecture in Figure 1. Multi-source treebanks are generated by concatenating treebanks from multiple source languages. We use the *arc-standard* algorithm (Nivre 2004) for parsing.[3] For each configuration (parsing state), we extract features using the same feature templates defined by Guo et al. (2015), which can be divided into two categories: (partial) lexical features and non-lexical features. The (partial) lexical features, i.e. the multilingual word/cluster features are projected to the embedding layer through the induced unified word representations. Note that the unified word embedding matrix is fixed during training, whereas all of the other feature embeddings get updated.

All features are then fed as input to the neural network classifier with the same structure as Chen and Manning (2014). We use the cross-entropy loss as objection function, and use mini-batch AdaGrad to train the parser.

# Experiments

This section describes the experiments. We first describe data and tools used in the experiments, and then the results.

## Data and Settings

**Data** We use the Google universal treebanks (v2.0) (McDonald et al. 2013) for evaluation. The languages we consider include all Indo-European languages presented in the universal treebanks. For both MULTI-SG and MULTI-PROJ, we use the Europarl corpus for EN-{DE, ES, FR, PT, IT, SV} parallel data,[4] and the WMT-2011 English news corpora as additional monolingual data.[5]

**Tools and Settings** We use the *cdec* (Dyer et al. 2010) alignment tool to obtain word alignments. We use the *word2vec* to train the seeding English word embeddings in

[3]For more details of the *arc-standard* algorithm, we refer the readers to Nivre (2004).

[4]http://www.statmt.org/europarl/

[5]http://www.statmt.org/wmt11/

| | MULTI-DELEX | | MULTI-SG | | MULTI-PROJ | | Guo15 | | Zhang15 | | Søgaard15 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | UAS | LAS | UAS | LAS | UAS | LAS | UAS | LAS | UAS | LAS | UAS | LAS |
| DE | 59.35 | 49.82 | 61.70 | 54.16 | **65.01** | **55.91** | 60.35 | 51.54 | 62.5 | 54.1 | 56.56 | 48.82 |
| ES | 75.54 | 64.68 | 78.42 | 71.56 | **79.00** | **73.08** | 71.90 | 62.28 | 78.0 | 68.3 | 64.03 | 55.03 |
| FR | 74.41 | 64.21 | 76.44 | 70.21 | **77.69** | **71.00** | 72.93 | 63.12 | **78.9** | 68.8 | 66.22 | 56.76 |
| IT | 76.60 | 65.49 | 77.48 | 70.04 | **78.49** | **71.24** | – | – | **79.3** | 69.4 | – | – |
| PT | 75.64 | 69.66 | 77.87 | 74.10 | **81.86** | **78.60** | – | – | 78.6 | 72.5 | – | – |
| SV | 73.38 | 62.90 | 76.45 | 67.74 | **78.28** | **69.53** | – | – | 75.0 | 62.5 | 67.32 | 57.70 |
| *AVG* | *72.49* | *62.79* | *74.73* | *67.97* | ***76.39*** | ***69.32*** | – | – | *75.4* | *65.9* | – | – |

Table 1: Parsing accuracies of different transfer approaches with cluster features on the test data using gold standard POS tags. All results are evaluated using both the unlabeled attachment score (UAS) and the labeled attachment score (LAS).

the *Multi-Proj* approach. For all parsing experiments, we adopt the implementation of Guo et al. (2015), and follow their hyper-parameter settings.

For word clustering, we use the multi-threaded Brown clustering tool to learn monolingual (EN) word clusters, and then projected to each of the rest languages independently.[6] The number of clusters are set to 256.

Considering that in practice when we apply our model to a low-resource language, typically we don't have any development data for parameter tuning. So we simply train our parsing models for 20,000 iterations without early-stopping.

Following previous work of Guo et al. (2015) and Zhang and Barzilay (2015), we use gold standard part-of-speech (POS) tags for both training and testing.

**Baseline Systems** We compare our system with the following baseline transfer systems:

- Bilingual transfer (Guo15). We consider the best models presented in Guo et al. (2015). In their approach, only English is used as source language. Bilingual word embeddings and clusters are induced for filling the lexical features gap. They only report results on DE, ES and FR.
- Multi-source delexicalized transfer (MULTI-DELEX). This can be viewed as a special case of our framework when all lexical and partial lexical features are discarded.

Besides, we also compare with two recently proposed model transfer systems which are closely related to ours, but using different parsers and resources.

- Hierarchical low-rank tensor model (Zhang15). Zhang and Barzilay (2015) proposed a hierarchical low-rank tensor model for multilingual transfer of dependency parsers utilizing the idea of selective parameter sharing.
- Inverted indexing (Søgaard15). Søgaard et al. (2015) obtained multilingual word embeddings based on inverted indexing of Wikipedia and applied them *mate-tools* for multi-source transfer parsing.

## Results

Table 1 summarizes the experimental results of various approaches on the test data.

[6] github.com/percyliang/brown-cluster

**Impact of Multi-Source** First, we demonstrate the effectiveness of exploiting multi-source treebanks for transfer by comparing MULTI-DELEX with Guo15. Even without any lexical features, MULTI-DELEX significantly outperforms the best bilingual transfer models in ES and FR.

**Impact of Unified Word Representation** By incorporating the unified lexical features (words, clusters) either by MULTI-SG or MULTI-PROJ, we observe large improvements against MULTI-DELEX. Respectively, MULTI-SG improves upon the delexicalized transfer models by an average of 2.24% of UAS and 5.18% of LAS. MULTI-PROJ obtains larger gains by an average of 3.90% of UAS and 6.53% of LAS. We can see that the LAS gains are much more significant than the UAS gains, which indicates that the lexical features indeed have larger impacts on labeled parsing than unlabeled parsing.

Overall the offline approach (MULTI-PROJ) works slightly better than the joint learning approach (MULTI-SG), which goes against our intuition. One reason might be that the joint learning approach cannot do well when the monolingual objective and bilingual objection do not agree in specific samples. We leave further analysis to future work.

Furthermore, we outperform the state-of-the-art transfer system Zhang15 (Zhang and Barzilay 2015) by an average of 0.99% in UAS and 3.42% in LAS. Their UAS of FR and IT are slightly higher than ours. One possible reason is the use of linguistic typological features in their model, which should be beneficial for two closely-related languages like FR and IT. Søgaard15 (Søgaard et al. 2015) only reported results on DE, ES, FR, and SV. However, their results are less promising.

## Target Language Adaptation with Minimal Supervision

This section investigates a more practical scenario, where minimal supervision is available for the target language of interest. Given the universal dependency grammar and a new language, it is not difficult to manually annotate dependency structures for a small amount (e.g. 50 sentences) of sentences. The same setting has also been explored in Zhang and Barzilay (2015) as a semi-supervised transfer scenario.

We follow Zhang and Barzilay (2015) for sampling 50 annotated sentences from target languages. Instead of combining the target language sentences with the multi-source

| | MULTI-DELEX(50) | | MULTI-SG(50) | | MULTI-PROJ(50) | | Zhang15(Semi) | |
|---|---|---|---|---|---|---|---|---|
| | UAS | LAS | UAS | LAS | UAS | LAS | UAS | LAS |
| DE | $67.26_{(+7.81)}$ | $57.40_{(+7.58)}$ | $72.76_{\mathbf{(+11.06)}}$ | $66.28_{\mathbf{(+12.12)}}$ | $\mathbf{73.61}_{\mathbf{(+8.60)}}$ | $\mathbf{66.79}_{\mathbf{(+10.88)}}$ | **74.2** | 65.6 |
| ES | $73.46_{(-2.08)}$ | $64.19_{(-0.49)}$ | $79.07_{(+0.65)}$ | $74.20_{(+2.64)}$ | $\mathbf{79.67}_{(+0.67)}$ | $\mathbf{74.27}_{(+1.19)}$ | 78.4 | 70.5 |
| FR | $74.60_{(+0.19)}$ | $64.72_{(+0.51)}$ | $79.26_{(+2.82)}$ | $73.10_{(+2.89)}$ | $\mathbf{79.99}_{(+2.30)}$ | $\mathbf{74.45}_{(+3.45)}$ | 79.6 | 71.8 |
| IT | $75.68_{(-0.92)}$ | $67.56_{(+2.07)}$ | $\mathbf{79.92}_{(+1.43)}$ | $74.86_{(+4.82)}$ | $79.85_{(+0.46)}$ | $\mathbf{74.94}_{(+3.70)}$ | **80.9** | 72.6 |
| PT | $75.01_{(-0.63)}$ | $68.37_{(-1.29)}$ | $\mathbf{81.44}_{(+3.57)}$ | $\mathbf{78.77}_{(+4.67)}$ | $81.11_{(-0.75)}$ | $78.22_{(-0.38)}$ | 79.3 | 73.5 |
| SV | $74.93_{(+1.55)}$ | $65.16_{(+2.26)}$ | $\mathbf{80.04}_{(+3.59)}$ | $\mathbf{74.10}_{(+6.36)}$ | $80.03_{(+1.75)}$ | $73.71_{(+4.18)}$ | 78.3 | 67.9 |
| *AVG* | $\mathit{73.49}_{(+1.00)}$ | $\mathit{64.57}_{(+1.78)}$ | $\mathit{78.75}_{(+4.02)}$ | $\mathit{73.55}_{(+5.58)}$ | $\mathbf{\mathit{79.04}}_{(+2.65)}$ | $\mathbf{\mathit{73.73}}_{(+4.41)}$ | *78.5* | *70.3* |

Table 2: Parsing accuracies of different transfer models with 50 annotated sentences from target languages as minimal supervision. The numbers in parentheses are absolute improvements over the directly transferred models as shown in Table 1.

treebanks to retrain the models, we consider an online strategy, by directly fine-tuning our transferred models using the sampled sentences with 100 iterations. Results are shown in Table 2. We observe significant improvements for both MULTI-SG and MULTI-PROJ against direct transfer.

Interestingly, all models on DE are dramatically improved with minimal supervision, inferring that DE has the most divergent syntactic structures with the source languages, which cannot be well recovered through cross-lingual transfer. To verify this, we investigate the *dobj* relation, which distinguish DE from other languages due to the V2 position of verbs. Table 3 demonstrates the significant distribution divergence between left-directed and right-directed *dobj* relation from the training data. We further examine the precision/recall improvement of *dobj* brought by minimal supervision. Table 4 verifies our assumption. As we can see, the recalls of *dobj* are improved dramatically and consistently.

| | | $dobj_{\curvearrowright}$ | $dobj_{\curvearrowleft}$ | ratio |
|---|---|---|---|---|
| Target | DE | 4,277 | 3,457 | 1.2 : 1 |
| Source | EN | 38,395 | 764 | 50.3 : 1 |
| | ES | 10,551 | 1,175 | 9.0 : 1 |
| | FR | 10,015 | 2,667 | 3.8 : 1 |
| | IT | 4,714 | 695 | 6.8 : 1 |
| | PT | 8,052 | 773 | 10.4 : 1 |
| | SV | 2,724 | 163 | 16.7 : 1 |

Table 3: Distribution divergence of left-directed and right-directed arcs with *dobj* relation across different languages.

| | MULTI-DELEX | | MULTI-SG | | MULTI-PROJ | |
|---|---|---|---|---|---|---|
| | P | R | P | R | P | R |
| Unsup | 36.84 | 35.69 | 36.10 | 38.65 | 50.47 | 35.69 |
| +50 | 39.62 | 41.45 | 47.38 | 60.86 | 52.34 | 62.66 |
| $\Delta$ | ***2.78*** | ***5.76*** | ***11.28*** | ***22.21*** | ***1.87*** | ***26.97*** |

Table 4: Effect of minimal supervision on *dobj* of DE. Unsup indicates the (unsupervised) directly transfer models.

Furthermore, we outperform the semi-supervised results of Zhang15(Semi) under the same setting by an averaged LAS of 3.53% (73.73 *vs.* 70.3).

## Related Work

There has been extensive research on annotation projection for cross-lingual parsing. A lot of work along this line has been dedicated to the process of robust projection, involving various innovations such as posterior regularization (Ganchev, Gillenwater, and Taskar 2009), entropy regularization and parallel guidance (Ma and Xia 2014), treebank translation (Tiedemann and Nivre 2014), and a most recent density-driven method (Rasooli and Collins 2015).

For model transfer, some additional works that are related to this study but under different settings include learning projection features (Durrett, Pauls, and Klein 2012) and utilizing typological features for selective sharing (Naseem, Barzilay, and Globerson 2012).

Overall, these two categories of methods are complementary and can be integrated to push the performance further.

## Conclusion

We propose a novel representation learning framework for multi-source transfer parsing. We introduced two algorithms for learning unified multilingual word representations to bridge the lexical feature gaps across multiple languages.

Experiments on Google universal dependency treebanks (v2.0) demonstrate the effectiveness of our framework. Our multi-source delexicalized model significantly outperform the strongest bilingual transfer model for most of the languages. By incorporating the unified word representations, our best models obtain large improvements against the delexicalized models by an average of 3.90% in UAS and 6.53% in LAS. We also outperform the state-of-the-art model transfer system (Zhang and Barzilay 2015).

We further investigate the effect of minimal supervision from target languages. We show that with only 50 annotated sentences, the model can be improved further, which is of great significance for practical scenario.

## Acknowledgments

We are grateful to Yuan Zhang for graciously providing the sampled data used in our minimal supervision experiment. We also thank the anonymous reviewers for the insightful comments and suggestions. This work was supported by the National Key Basic Research Program of China via grant 2014CB340503 and the National Natural Science Foundation of China (NSFC) via grant 61133012 and 61370164.

## References


Chandar, S.; Lauly, S.; Larochelle, H.; Khapra, M.; Ravindran, B.; Raykar, V.; and Saha, A. 2014. An autoencoder approach to learning bilingual word representations. In *NIPS*, 1853–1861.

Chen, D., and Manning, C. D. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*, 740–750.

Duong, L.; Cohn, T.; Bird, S.; and Cook, P. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *ACL-IJCNLP*, 845–850.

Durrett, G.; Pauls, A.; and Klein, D. 2012. Syntactic transfer using a bilingual lexicon. In *EMNLP*, 1–11.

Dyer, C.; Lopez, A.; Ganitkevitch, J.; Weese, J.; Ture, F.; Blunsom, P.; Setiawan, H.; Eidelman, V.; and Resnik, P. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *ACL*.

Dyer, C.; Ballesteros, M.; Ling, W.; Matthews, A.; and Smith, N. A. 2015. Transition-based dependency parsing with stack long short-term memory. In *ACL-IJCNLP*, 334–343.

Faruqui, M., and Dyer, C. 2014. Improving vector space word representations using multilingual correlation. In *EACL*, 462–471.

Firth, J. R. 1957. A synopsis of linguistic theory 19301955. In *In Studies in linguistic analysis*, 1–32.

Ganchev, K.; Gillenwater, J.; and Taskar, B. 2009. Dependency grammar induction via bitext projection constraints. In *ACL-IJCNLP*, 369–377.

Gouws, S.; Bengio, Y.; and Corrado, G. 2014. Bilbowa: Fast bilingual distributed representations without word alignments. *arXiv preprint arXiv:1410.2455*.

Guo, J.; Che, W.; Yarowsky, D.; Wang, H.; and Liu, T. 2015. Cross-lingual dependency parsing based on distributed representations. In *ACL-IJCNLP*, 1234–1244.

Hermann, K. M., and Blunsom, P. 2014. Multilingual models for compositional distributed semantics. In *ACL*, 58–68.

Hwa, R.; Resnik, P.; Weinberg, A.; Cabezas, C.; and Kolak, O. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering* 11(03):311–325.

Klein, D., and Manning, C. D. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *ACL*, 478.

Klementiev, A.; Titov, I.; and Bhattarai, B. 2012. Inducing crosslingual distributed representations of words. In *COLING*, 1459–1474.

Lu, A.; Wang, W.; Bansal, M.; Gimpel, K.; and Livescu, K. 2015. Deep multilingual correlation for improved word embeddings. In *NAACL*.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Bilingual word representations with monolingual quality in mind. In *NAACL Workshop on Vector Space Models*, 151–159.

Ma, X., and Xia, F. 2014. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *ACL*, 1337–1348.

McDonald, R., and Nivre, J. 2007. Characterizing the errors of data-driven dependency parsing models. In *EMNLP-CoNLL*, 122–131.

McDonald, R. T.; Nivre, J.; Quirmbach-Brundage, Y.; Goldberg, Y.; Das, D.; Ganchev, K.; Hall, K. B.; Petrov, S.; Zhang, H.; Täckström, O.; et al. 2013. Universal dependency annotation for multilingual parsing. In *ACL*, 92–97.

McDonald, R.; Petrov, S.; and Hall, K. 2011. Multi-source transfer of delexicalized dependency parsers. In *EMNLP*, 62–72.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Naseem, T.; Barzilay, R.; and Globerson, A. 2012. Selective sharing for multilingual dependency parsing. In *ACL*, 629–637.

Nivre, J. 2004. Incrementality in deterministic dependency parsing. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*, 50–57.

Rasooli, M. S., and Collins, M. 2015. Density-driven cross-lingual transfer of dependency parsers. In *EMNLP*, 328–338.

Søgaard, A.; Agić, v.; Martínez Alonso, H.; Plank, B.; Bohnet, B.; and Johannsen, A. 2015. Inverted indexing for cross-lingual nlp. In *ACL-IJCNLP*, 1713–1722.

Täckström, O.; McDonald, R.; and Uszkoreit, J. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *NAACL*, 477–487.

Tiedemann, J., and Nivre, J. 2014. Treebank translation for cross-lingual parser induction. *CoNLL* 130.

Tiedemann, J. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *COLING*.

Weiss, D.; Alberti, C.; Collins, M.; and Petrov, S. 2015. Structured training for neural network transition-based parsing. In *ACL-IJCNLP*, 323–333.

Xiao, M., and Guo, Y. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *CoNLL*, 119–129.

Zhang, Y., and Barzilay, R. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *EMNLP*, 1857–1867.

Zhang, Y., and Nivre, J. 2011. Transition-based dependency parsing with rich non-local features. In *ACL*, 188–193.

Zhou, H.; Zhang, Y.; Huang, S.; and Chen, J. 2015. A neural probabilistic structured-prediction model for transition-based dependency parsing. In *ACL-IJCNLP*, 1213–1222.

Zou, W. Y.; Socher, R.; Cer, D.; and Manning, C. D. 2013. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, 1393–1398.