# Cross-lingual Transfer of Named Entity Recognizers without Parallel Corpora

**Ayah Zirikly***
Department of Computer Science
The George Washington University
Washington DC, USA
ayaz@gwu.edu

**Masato Hagiwara**
Duolingo, Inc.
Pittsburgh PA, USA
masato@duolingo.com

## Abstract

We propose an approach to cross-lingual named entity recognition model transfer without the use of parallel corpora. In addition to global de-lexicalized features, we introduce multilingual gazetteers that are generated using graph propagation, and cross-lingual word representation mappings without the use of parallel data. We target the e-commerce domain, which is challenging due to its unstructured and noisy nature. The experiments have shown that our approaches beat the strong MT baseline, where the English model is transferred to two languages: Spanish and Chinese.

## 1 Introduction

Named Entity Recognition (NER) is usually solved by a supervised learning approach, where sequential labeling models are trained from a large amount of manually annotated corpora. However, such rich annotated data only exist for resource-rich languages such as English, and building NER systems for the majority of resource-poor languages, or specific domains in *any* languages, still poses a great challenge.

Annotation projection through parallel text (Yarowsky et al., 2001), (Das and Petrov, 2011), (Wang and Manning, 2014) has been traditionally used to overcome this issue, where the annotated tags in the source (resource-rich) language are projected via word-aligned bilingual parallel text (bitext) and used to train sequential labeling models in the (resource-poor) target language. However, this could lead to two issues: firstly, word alignment and projected tags are potentially noisy, making the trained models sub-optimal. Instead of projecting noisy labels explicitly, Wang and Manning (2014) project posterior marginals expectations as soft constraints. Das and Petrov (2011) projected POS tags from source language types to target language trigarms using graph propagation and used the projected label distribution to train robust POS taggers. Secondly, the availability of such bitext is limited especially for resource-poor languages and domains, where it is often the case that available resources are moderately-sized monolingual/comparable corpora and small bilingual dictionaries.

Instead, we seek a *direct transfer* approach (Figure 1) to cross-lingual NER (also classified as transductive transfer learning (Pan and Yang, 2010) and closely related to domain adaptation). Specifically, we only assume the availability of *comparable* corpora and small-sized bilingual dictionaries, and use the same sequential tagging model trained on the source corpus for tagging the target corpus. Direct transfer approaches are extensively studied for cross-lingual dependency parser transfer. For example, Zeman et al. (2008) built a constituent parser using direct transfer between closely related languages, namely, Danish and Swedish. McDonald et al. (2011) trained de-lexicalized dependency parsers in English and then "re-lexicalized" the parser. However, cross-lingual transfer of named entity taggers have not been studied enough, and this paper, to the best of the authors' knowledge, is the first to apply direct transfer learning to NER.

Transfer of NER taggers poses a difficult challenge that is different from syntax transfer: most of the past work deals with de-lexicalized parsers, yet one of the most important clues for NER, gazetteers, is inherently lexicalized. Also, various features used for dependency parsing (Universal POS tags, unsupervised clustering, etc.) are yet to

[*] This work has been performed while the authors were at Rakuten Institute of Technology, New York. The authors would like to thank Prof. Satoshi Sekine at New York University and other members of Rakuten Institute of Technology for their support during the project.

*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*
*and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 390–396,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

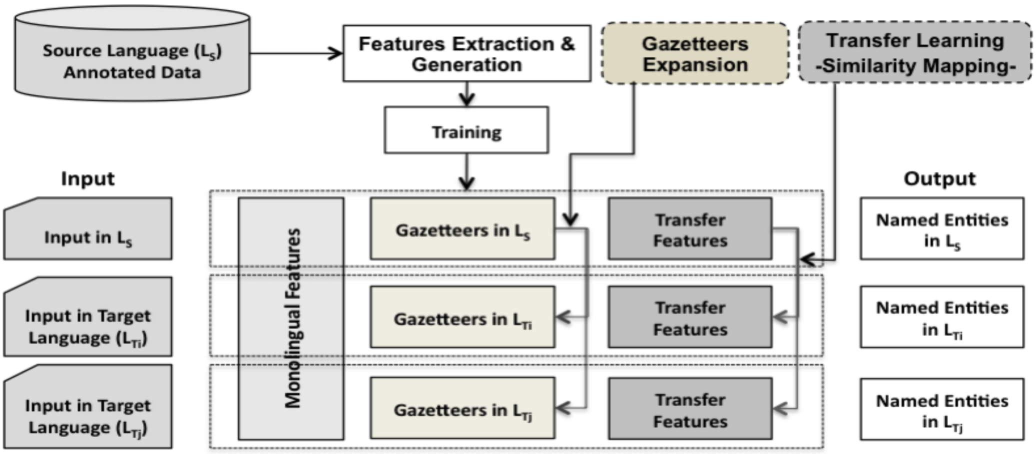
Source Language ($L_S$) Annotated Data
Features Extraction & Generation
Gazetteers Expansion
Transfer Learning -Similarity Mapping-
Training
Input
Output
Input in $L_S$
Input in Target Language ($L_{Ti}$)
Input in Target Language ($L_{Tj}$)
Monolingual Features
Gazetteers in $L_S$
Gazetteers in $L_{Ti}$
Gazetteers in $L_{Tj}$
Transfer Features
Transfer Features
Transfer Features
Named Entities in $L_S$
Named Entities in $L_{Ti}$
Named Entities in $L_{Tj}$

Figure 1: System Framework

be proven useful for direct transfer of NER. Therefore, the contributions of this paper is as follows:

1. We show that direct transfer approach for multilingual NER actually works and performs better than the strong MT baseline (Shah et al., 2010), where the system's output in the source language is simply machine translated into the target language.

2. We explore various non-lexical features, namely, Universal POS tags and Brown cluster mapping, which are deemed effective for multilingual NER transfer. Although brown cluster mapping (Täckström et al., 2012), Universal POS Tagset (Petrov et al., 2011), and re-lexicalization and self training (Täckström et al., 2013) are shown to be effective for direct transfer of dependency parsers, there have been no studies exploring these features for NER transfer.

3. We show that gazetteers can actually be generated only from the source language gazetteers and a comparable corpus, through a technique which we call *gazetteer expansion* based on semi-supervised graph propagation (Zhu et al., 2003). Gazetteer expansion has been used for various other purposes, including POS tagging (Alexandrescu and Kirchhoff, 2007) and dependency parsers (Durrett et al., 2012).

## 2 Approach

In this paper we propose a direct transfer learning approach to train NER taggers in a multilingual setting. Our goal is to identify named entities in a target language $L_T$, given solely annotated data in the source language $L_S$. Previous approaches rely on parallel data to transfer the knowledge from one language to another. However, parallel data is very expensive to construct and not available for all language pairs in all domains. Thus, our approach loosens the constraint and only requires in-domain comparable corpora.

### 2.1 Monolingual NER in Source Language

Our framework is based on *direct transfer* approach, where we extract abstract, language-independent and non lexical features $F_S$ and $F_T$ in $L_S$ and $L_T$. A subset of $F_T$ is generated using a mapping scheme discussed in Section 2.2, then, directly apply $L_S$ NER model on $L_T$ using $F_T$. We adopt Conditional Random Field (CRF) sequence labeling (Lafferty et al., 2001) to train our system and generate the English model.

**Monolingual Features** 1) *Token position*: Instead of using token exact position, we use token relative position in addition to position's binary features such as token is in: first, second, and last third of the sentence. These features are based on the observation that certain tokens, such as brand names in title or description of a product, tend to appear at the beginning of the sentence, while others toward the end.

2) *Word Shape*: We use a list of binary features: is-alphanumerical, is-number, is-alpha, is-punctuation, the number length (if is-num is true), pattern-based features (e.g. regular expressions to capture certain patterns such as products model numbers), latin-only features (first-is-capital, all-capitals, all-small);

3) *In-Title*: A binary feature that specifies whether the token is in the product's title or description. For instance, brand names mostly appear in the beginning of titles, while this does not hold in descriptions;

4) *Preceding/Proceeding keywords within window*: some NEs are often preceded by certain keywords. For instance, often a product size is preceded by certain keywords such as dimension, height or word"size." In our work we use a manually created list of keywords for two classes Color and Size. Although the keyword list is domain dependent, it is often short and can be easily updated.

5) *Universal Part of Speech Tags*: Part of Speech (POS) tags have been widely used in many NER systems. However, each language has its own POS tagset that often has limited overlap with other POS languages' tagsets. Thus, we use a coarse-grained layer of POS tags called Universal POS, as proposed in (Petrov et al., 2011).

6) *Token is a unit*: A binary feature that is set to true if it matches an entry in the units dictionary (e.g., "cm.")

7) *Gazetteers*: Building dictionaries for every $L_T$ of interest is expensive; thus, we propose a method, described in Section 3, to generate gazetteers in $L_T$ given ones in $L_S$.

8) *Brown Clustering (BC)*: Word representations, especially Brown Clustering (Brown et al., 1992), are used in many NLP tasks and are proven to improve NER performance (Turian et al., 2010). In this work, we use cluster IDs of variable prefix lengths in order to retrieve word similarities on different granularity levels.

## 2.2 Multilingual NER in Target Language

Our goal is to transfer each feature from $L_S$ to $L_T$ space. The main challenge resides in transferring features 7 and 8 without the use of external resources and parallel data for every target language.

### 2.2.1 Brown Clustering Mapping

Given i) Vocabulary in the source/target languages $V_S = \{w_1^S, w_2^S, ..., w_{N_S}^S\}$ and $V_T = \{w_1^T, w_2^T, ..., v_{N_T}^T\}$; ii) The output of brown clustering on $L_S$ and $L_T$: $C_S = \{c_1^S, ..., c_{K_S}^S\}$ and $C_T = \{c_1^T, ..., c_{K_L}^T\}$, we aim to find the best mapping $c^{S*}$ that maximizes the cluster similarity $sim_C$ for each target cluster (Equation 1), and for each metric discussed in the following. We calculate the cluster similarity $sim_C$ as the weighted average of the word similarity $sim_W$ of the members of the two clusters (Equation 2).

$$c^{S*} = \arg\max_{c^S \in C_S} sim_C(c^S, c^T) \text{ for each } c_T \in C_T \quad (1)$$

$$sim_C(c_t, c_s) = \frac{1}{|c^S||c^T|} \sum_{w^S \in c^S, w^T \in c^T} sim_W(w^S, w^T) \quad (2)$$

**Clusters Similarity Metrics** The similarity metrics used can be summarized in:

a) *String Similarity* (external resources independent): This metric works only on languages that share the same alphabet, as it is based on the intuition that most NEs conserve the name's shape or present minor changes that can be identified using edit distance in closely related languages (we use Levenshtein distance (Levenshtein, 1966)). The two variations of string similarity metrics used are: i) *Exact match*: $sim_W(w_i, w_j) = 1$ if $w_i = w_j$; ii) *Edit distance*: $sim_W(w_i, w_j) = 1$ if levenshtein-distance$(w_i, w_j) < \theta$.

b) *Dictionary-based similarity*: We present two similarity metrics using BabelNet synsets (Navigli and Ponzetto, 2012): i) *Binary co-occurence*: $sim_W^{binary}(w_i, w_j) = 1$ if $w_j \in synset(w_i)$, where $synset(w_i)$ is the set of words in the BabelNet synset of $w_i$; ii) *Frequency weighted*: Weighted version of the binary similarity that is based on the observation that less frequent words tend to be less reliable in brown clustering: $sim_W^{weighted}(w_i, w_j) = [\log f(w_i) + \log f(w_j)] \times sim_W^{binary}(w_i, w_j)$ where $f(w)$ is the frequency of word $w$. Unlike String similarity metrics, this metric is not limited to similar languages due to the use of multilingual dictionaries i.e., BabelNet, which covers 271 languages.

# 3 Gazetteer expansion

In our approach, we use graph-based semi-supervised learning to expand the gazetteers in the source language to the target. Figure 2 illustrates the motivation of our approach. Suppose we have "New York" in the GPE gazetteer in $L_S$ (English in this case), and we would like to bootstrap the corresponding GPE gazetteer in $L_T$ (Spanish). Although there is no direct link between "New York" and "Nueva York," you can infer that "Puerto Rico" (in English) is similar to "New York" based on some intra-language semantic similarity model, then "Puerto Rico" is actually

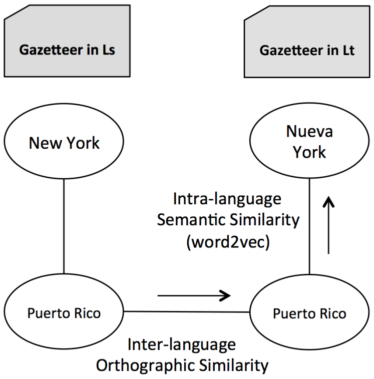
Gazetteer in Ls
Gazetteer in Lt
New York
Nueva York
Intra-language Semantic Similarity (word2vec)
Puerto Rico
Puerto Rico
Inter-language Orthographic Similarity

Figure 2: Gazeteer expansion

identical in both languages, then finally "Nueva York" is similar to "Puerto Rico" (in Spanish) again based on the Spanish intra-language similarity model. This indirect inference of beliefs from the source gazetteers to the target can be modeled by semi-supervised graph propagation (Zhu et al. 2003), where graph nodes are $V_S \cup V_T$, positive labels are entries in the $L_S$ gazetteer (e.g., GPE) which we wish to expand to $L_T$, and negative labels are entries in other gazetteers (e.g., PERSON) in $L_S$. The edge weights between same-language nodes $w_i$ and $w_j$ are given by $\exp(-\sigma||\mathbf{w}_i - \mathbf{w}_j||)$ where $\mathbf{w}_i$ is the distributed vector representation of word $w_i$ computed by word2vec (Mikolov et al., 2013). The edge weights between node $w_i \in V_S$ and $v_j \in V_T$ are defined 1 if the spelling of these two words are identical and 0 otherwise. Note that this spelling based similarity propagation is still available for language pairs with different writing systems such as English and Chinese, because major NEs (e.g., brand names) are often written in Roman alphabets even in Chinese products. Since the analytical solution to this propagation involves the computation of $n \times n$ ($n$ is the number of unlabeled nodes) matrix, we approximated it by running three propagation steps iteratively, namely, $L_S \rightarrow L_S$, $L_S \rightarrow L_T$, and $L_T \rightarrow L_T$. After the propagation, we used all the nodes with their propagated values $f(w_i) > \theta$ as entities in the new gazetteer.

## 4 Experiments

### 4.1 Datasets

The targeted dataset contains a list of products (titles and descriptions). The titles of products are $\approx$ 10 words long and poorly structured, adding more difficulties to our task. On the other hand, the length of product descriptions ranges from 12-130 words. The e-commerce genre poses the need to introduce new NE tagset as opposed to the conventional ones, thus we introduce 6 tag types: 1) Color; 2) Brand names; 3) Size; 4) Type: e.g. "camera," "shirt"; 5) Material: e.g. "plastic", "cotton"; 6) Model: the model number of a product: e.g., "A1533.". For the rest of the experiments, English (EN) is the source language, whereas we experiment with Spanish (ES) and Chinese (ZH) as target languages. The datasets used are: i) Training data: 1800 annotated English products from Rakuten.com shopping (Rakuten, 2013a); ii) Test data: 300 ES products from Rakuten Spain (Rakuten, 2013b) and 500 products from Rakuten Taiwan (Rakuten, 2013c); iii) Brown clustering: *English*: Rakuten shopping 2013 dump (19m unique products with 607m tokens); *Spanish*: Rakuten Spain 2013 dump (700K unique products that contains 41m tokens) in addition to Spanish Wikipedia dump (Al-Rfou', 2013); *Chinese*: Wikipedia Chinese 2014 dump (147m tokens) plus 16k products crawled from Rakuten Taiwan. Table 1 shows the numbers of tags per category for each language.

| | Color | Brand | Material | Model | Type | Size |
|---|---|---|---|---|---|---|
| EN | 358 | 814 | 733 | 203 | 1238 | 427 |
| ES | 207 | 425 | 301 | 172 | 606 | 126 |
| ZH | 416 | 60 | 381 | 24 | 690 | 306 |

Table 1: Language-Tags Numbers Stats

### 4.2 Baseline

To the best of our knowledge, there is no previous work that proposes transfer learning for NER without the use of parallel data. Thus, we ought to generate a strong baseline to compare our results to. Given the language pair $(L_S, L_T)$, we use Microsoft Bing Translate API to generate $L_T \rightarrow L_S$ translation. Then, we apply $L_S$ NER model on the translated text and evaluate by mapping the tagged tokens back to $L_T$ using the word alignments generated by Bing Translate. We choose Bing translate as opposed to Google translate due to its free-to-use API that provides word alignment information on the character level.

### 4.3 Results & Discussion

For each studied language we use Stanford CoreNLP (Manning et al., 2014) for EN and ZH, and TreeTagger (Schmid, 1994) for ES to produce

|  | Color | Brand | Material | Model | Type | Size | Micro-Avg |
|---|---|---|---|---|---|---|---|
| EN-Mono | 68.45 | 71.91 | 50.94 | 59.78 | 53.73 | 45.42 | 61.12 |
| ES-Baseline | 24.23 | 3.44 | 13.08 | 14.51 | 12.5 | 6.61 | 13.79 |
| ES-TL | 18.00 | 9.37 | 8.05 | 16.99 | 18.26 | 10.64 | **39.46** |
| ES-GT | 38.49 | 13.31 | 33.5 | 2.27 | 36.43 | 1.16 | 30.20 |
| ZH-Baseline | 19.16 | 2.79 | 11.96 | None | 9.35 | 6.34 | 12.58 |
| ZH-TL | 9.36 | 1.02 | 1.81 | None | 17.28 | 17.74 | **23.43** |

Table 2: F-score Results

the tokens and the POS tags. However, we apply extra processing steps to the tokenizer due to the nature of the domain's data (e.g., avoid tokenizing models instances), in addition to normalizing URLs, numbers, and elongation. We also map POS tags for all the source and target languages to the universal POS tagset as explained in 2.1.

Based on Table 2, we note that English monolingual performance (80:20 train/test split and 5-folds cross variation) is considerably lower than state-of-the-art English NER systems, which is due to the nature of our targeted domain, the newly proposed NE tagset, and most importantly, the considerably small training data (1280 products). These factors also affects the baseline and our proposed system performance.

Table 2 illustrates the results for the English monolingual NER system (EN-Mono), baseline for ES and ZH (ES-Baseline and ZH-Baseline, respectively), our proposed transfer learning approach with the gazetteer expansion (ES-TL and ZH-TL). Additionally, we added the results of our proposed approach where the gazetteers used are machine translated using Google translate from the English gazetteers to Spanish (ES-MT), in order to evaluate our gazetteer expansion approach performance to the translated gazetteers.

We note that ES-Baseline and ZH-Baseline are considerably low due to the poor word alignment generated by Bing Translator, which results in incorrect tag projection. The quality of mapping is mainly due to the noisy nature of the domain's data, which can be very expensive to fix.

Although the performance of our proposed system is low (39.46% for ES and 23.43% for ZH), but it surpasses the baseline performance in most of the tag classes and yields an overall improvement on the micro-average F-score of $\approx 23\%$ in ES and $11\%$ in ZH. We note that one of the reasons behind ZH *Brand* low performance is that universal-POS for brands in EN are mostly proper noun as opposed to noun in ZH, additionally the considerably low number of brands in ZH test data (60). On the other hand, it is intuitive that *Model* yields one of the best performance among the tags, since it is the most language independent tag (as depicted in ES-TL). However, this does not hold true in ZH due to the very small number of *Model* instances (24). *Type* produces the best performance in ES and ZH, due to the high coverage of the new expanded gazetteer over *Type* instances, in addition to the large number of training instances (1238), in comparison to the other tags. After conducting leave-out experiments on Brown clustering and gazetteers features in ES, we note that both shows an improvement of $\approx 4\%$ and $\approx 8\%$ respectively.

Our system surpasses the MT-based gazetter expansion by $\approx 9\%$, when comparing ES-TL to ES-MT. However, as expected the main improvement is in *Model* and *Size* tags as opposed to other tags (e.g. *Brand* and *Color*) where MT provides more accurate gazetteers. In our system output, colors that are included in $L_T$ expanded gazetteers (e.g. "azul" in ES) and have a high similarity score in our proposed BC mapping, are correctly tagged. On the other hand OOV Brand have a very large prediction error rate due to the small training data.

## 5 Conclusion and Future Works

In this paper, we propose a cross-lingual NER transfer learning approach which does not depend on parallel corpora. Our experiments showed the ability to transfer NER model to latin (ES) and non latin (ZH) languages. For the future work, we would like to investigate the generality of our approach in broader languages and domains.

## References

Rami Al-Rfou'. 2013. Spanish wikipedia dump. url = https://sites.google.com/site/rmyeid/projects/polyglot.

Andrei Alexandrescu and Katrin Kirchhoff. 2007. Data-driven graph construction for semi-supervised graph-based learning in NLP. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 204–211, Rochester, New York, April. Association for Computational Linguistics.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

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

Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1–11. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

VI Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, October.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Rakuten. 2013a. Rakuten shopping. url = http://www.rakuten.com/.

Rakuten. 2013b. Rakuten spanish.

Rakuten. 2013c. Rakuten taiwanese.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees.

Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. 2010. Synergy: A named entity recognition system for resource-scarce languages such as swahili using online machine translation. In *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, pages 21–26.

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

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071, Atlanta, Georgia, June. Association for Computational Linguistics.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

Mengqiu Wang and Christopher D Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning.

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

Daniel Zeman, Univerzita Karlova, and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *In IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *IN ICML*, pages 912–919.