Employing Word Representations and Regularization for Domain Adaptation of Relation Extraction

Thien Huu Nguyen
Computer Science Department
New York University
New York, NY 10003 USA
thien@cs.nyu.edu

Ralph Grishman
Computer Science Department
New York University
New York, NY 10003 USA
grishman@cs.nyu.edu

Abstract

Relation extraction suffers from a performance loss when a model is applied to out-of-domain data. This has fostered the development of domain adaptation techniques for relation extraction. This paper evaluates word embeddings and clustering on adapting feature-based relation extraction systems. We systematically explore various ways to apply word embeddings and show the best adaptation improvement by combining word cluster and word embedding information. Finally, we demonstrate the effectiveness of regularization for the adaptability of relation extractors.

1 Introduction

The goal of Relation Extraction (RE) is to detect and classify relation mentions between entity pairs into predefined relation types such as *Employment* or *Citizenship* relationships. Recent research in this area, whether feature-based (Kambhatla, 2004; Boschee et al., 2005; Zhou et al., 2005; Grishman et al., 2005; Jiang and Zhai, 2007a; Chan and Roth, 2010; Sun et al., 2011) or kernel-based (Zelenko et al., 2003; Bunescu and Mooney, 2005a; Bunescu and Mooney, 2005b; Zhang et al., 2006; Qian et al., 2008; Nguyen et al., 2009), attempts to improve the RE performance by enriching the feature sets from multiple sentence analyses and knowledge resources. The fundamental assumption of these supervised systems is that the training data and the data to which the systems are applied are sampled independently and identically from the same distribution. When there is a mismatch between data distributions, the RE performance of these systems tends to degrade dramatically (Plank and Moschitti, 2013). This is where we need to resort to domain adaptation techniques (DA) to adapt a model trained on one domain (the source domain) into a new model which can perform well on new domains (the target domains).

The consequences of linguistic variation between training and testing data on NLP tools have been studied extensively in the last couple of years for various NLP tasks such as Part-of-Speech tagging (Blitzer et al., 2006; Huang and Yates, 2010; Schnabel and Schütze, 2014), named entity recognition (Daumé III, 2007) and sentiment analysis (Blitzer et al., 2007; Daumé III, 2007; Daumé III et al., 2010; Blitzer et al., 2011), etc. Unfortunately, there is very little work on domain adaptation for RE. The only study explicitly targeting this problem so far is by Plank and Moschitti (2013) who find that the out-of-domain performance of kernel-based relation extractors can be improved by embedding semantic similarity information generated from word clustering and latent semantic analysis (LSA) into syntactic tree kernels. Although this idea is interesting, it suffers from two major limitations:

+ It does not incorporate word cluster information at different levels of granularity. In fact, Plank and Moschitti (2013) only use the 10-bit cluster prefix in their study. We will demonstrate later that the adaptability of relation extractors can benefit significantly from the addition of word cluster features at various granularities.

+ It is unclear if this approach can encode real-valued features of words (such as word embeddings (Mnih and Hinton, 2007; Collobert and Weston, 2008)) effectively. As the real-valued features are able to capture latent yet useful properties of words, the augmentation of lexical terms with these features is desirable to provide a more general representation, potentially helping relation extractors perform more robustly across domains.

In this work, we propose to avoid these limitations by applying a feature-based approach for RE which allows us to integrate various word features of generalization into a single system more natu-

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 68–74,
Baltimore, Maryland, USA, June 23-25 2014.

rally and effectively.

The application of word representations such as word clusters in domain adaptation of RE (Plank and Moschitti, 2013) is motivated by its successes in semi-supervised methods (Chan and Roth, 2010; Sun et al., 2011) where word representations help to reduce data-sparseness of lexical information in the training data. In DA terms, since the vocabularies of the source and target domains are usually different, word representations would mitigate the lexical sparsity by providing general features of words that are shared across domains, hence bridge the gap between domains. The underlying hypothesis here is that the absence of lexical target-domain features in the source domain can be compensated by these general features to improve RE performance on the target domains.

We extend this motivation by further evaluating word embeddings (Bengio et al., 2001; Bengio et al., 2003; Mnih and Hinton, 2007; Collobert and Weston, 2008; Turian et al., 2010) on feature-based methods to adapt RE systems to new domains. We explore the embedding-based features in a principled way and demonstrate that word embedding itself is also an effective representation for domain adaptation of RE. More importantly, we show empirically that word embeddings and word clusters capture different information and their combination would further improve the adaptability of relation extractors.

2 Regularization

Given the more general representations provided by word representations above, how can we learn a relation extractor from the labeled source domain data that generalizes well to new domains? In traditional machine learning where the challenge is to utilize the training data to make predictions on unseen data points (generated from the same distribution as the training data), the classifier with a good generalization performance is the one that not only fits the training data, but also avoids ovefitting over it. This is often obtained via regularization methods to penalize complexity of classifiers. Exploiting the shared interest in generalization performance with traditional machine learning, in domain adaptation for RE, we would prefer the relation extractor that fits the source domain data, but also circumvents the overfitting problem over this source domain[1] so that it could generalize well on new domains. Eventually, regularization methods can be considered naturally as a simple yet general technique to cope with DA problems.

Following Plank and Moschitti (2013), we assume that we only have labeled data in a single source domain but no labeled as well as unlabeled target data. Moreover, we consider the single-system DA setting where we construct a single system able to work robustly with different but related domains (multiple target domains). This setting differs from most previous studies (Blitzer et al., 2006) on DA which have attempted to design a specialized system for every specific target domain. In our view, although this setting is more challenging, it is more practical for RE. In fact, this setting can benefit considerably from our general approach of applying word representations and regularization. Finally, due to this setting, the best way to set up the regularization parameter is to impose the same regularization parameter on every feature rather than a skewed regularization (Jiang and Zhai, 2007b).

3 Related Work

Although word embeddings have been successfully employed in many NLP tasks (Collobert and Weston, 2008; Turian et al., 2010; Maas and Ng, 2010), the application of word embeddings in RE is very recent. Kuksa et al. (2010) propose an abstraction-augmented string kernel for bio-relation extraction via word embeddings. In the surge of deep learning, Socher et al. (2012) and Khashabi (2013) use pre-trained word embeddings as input for Matrix-Vector Recursive Neural Networks (MV-RNN) to learn compositional structures for RE. However, none of these works evaluate word embeddings for domain adaptation of RE which is our main focus in this paper.

Regarding domain adaptation, in representation learning, Blitzer et al. (2006) propose structural correspondence learning (SCL) while Huang and Yates (2010) attempt to learn a multi-dimensional feature representation. Unfortunately, these methods require unlabeled target domain data which are unavailable in our single-system setting of DA. Daumé III (2007) proposes an easy adaptation framework (EA) which is later extended to a semi-supervised version (EA++) to incorporate unla-

[1] *domain overfitting* (Jiang and Zhai, 2007b)

beled data (Daumé III et al., 2010). In terms of word embeddings for DA, recently, Xiao and Guo (2013) present a log-bilinear language adaptation framework for sequential labeling tasks. However, these methods assume some labeled data in target domains and are thus not applicable in our setting of unsupervised DA. Above all, we move one step further by evaluating the effectiveness of word embeddings on domain adaptation for RE which is very different from the principal topic of sequence labeling in the previous research.

4 Word Representations

We consider two types of word representations and use them as additional features in our DA system, namely Brown word clustering (Brown et al., 1992) and word embeddings (Bengio et al., 2001). While word clusters can be recognized as an one-hot vector representation over a small vocabulary, word embeddings are dense, low-dimensional, and real-valued vectors (distributed representations). Each dimension of the word embeddings expresses a latent feature of the words, hopefully reflecting useful semantic and syntactic regularities (Turian et al., 2010). We investigate word embeddings induced by two typical language models: Collobert and Weston (2008) embeddings (C&W) (Collobert and Weston, 2008; Turian et al., 2010) and Hierarchical log-bilinear embeddings (HLBL) (Mnih and Hinton, 2007; Mnih and Hinton, 2009; Turian et al., 2010).

5 Feature Set

5.1 Baseline Feature Set

Sun et al. (2011) utilize the full feature set from (Zhou et al., 2005) plus some additional features and achieve the state-of-the-art feature-based RE system. Unfortunately, this feature set includes the *human-annotated* (gold-standard) information on entity and mention types which is often missing or noisy in reality (Plank and Moschitti, 2013). This issue becomes more serious in our setting of single-system DA where we have a single source domain with multiple dissimilar target domains and an automatic system able to recognize entity and mention types very well in different domains may not be available. Therefore, following the settings of Plank and Moschitti (2013), we will only assume entity boundaries and not rely on the gold standard information in the experiments. We apply the same feature set as Sun et al. (2011) but remove the entity and mention type information[2].

5.2 Lexical Feature Augmentation

While Sun et al. (2011) show that adding word clusters to the heads of the two mentions is the most effective way to improve the generalization accuracy, the right lexical features into which word embeddings should be introduced to obtain the best adaptability improvement are unexplored. Also, which dimensionality of which word embedding should we use with which lexical features? In order to answer these questions, following Sun et al. (2011), we first group lexical features into 4 groups and rank their importance based on linguistic intuition and illustrations of the contributions of different lexical features from various feature-based RE systems. After that, we evaluate the effectiveness of these lexical feature groups for word embedding augmentation individually and incrementally according to the rank of importance. For each of these group combinations, we assess the system performance with different numbers of dimensions for both C&W and HLBL word embeddings. Let M1 and M2 be the first and second mentions in the relation. Table 1 describes the lexical feature groups.

Rank	Group	Lexical Features
1	**HM**	HM1 (head of M1)
		HM2 (head of M2)
2	**BagWM**	WM1 (words in M1)
		WM2 (words in M2)
3	**HC**	heads of chunks in context
4	**BagWC**	words of context

Table 1: Lexical feature groups ordered by importance.

6 Experiments

6.1 Tools and Data

Our relation extraction system is hierarchical (Bunescu and Mooney, 2005b; Sun et al., 2011) and apply maximum entropy (MaxEnt) in the MALLET[3] toolkit as the machine learning tool. For Brown word clusters, we directly apply the clustering trained by Plank and Moschitti (2013)

[2]We have the same observation as Plank and Moschitti (2013) that when the gold-standard labels are used, the impact of word representations is limited since the gold-standard information seems to dominate. However, whenever the gold labels are not available or inaccurate, the word representations would be useful for improving adaptability performance. Moreover, in all the cases, regularization methods are still effective for domain adaptation of RE.

[3]http://mallet.cs.umass.edu/

	System	In-domain (bn+nw)					Out-of-domain (bc development set)				
		C&W,25	C&W,50	C&W,100	HLBL,50	HLBL,100	C&W,25	C&W,50	C&W,100	HLBL,50	HLBL,100
1	Baseline	51.4	51.4	51.4	51.4	51.4	49.0	49.0	49.0	49.0	49.0
2	1+HM_ED	54.0(+2.6)	54.1(+2.7)	**55.7(+4.3)**	53.7(+2.3)	55.2(+3.8)	51.5(+2.5)	**52.7(+3.7)**	52.5(+3.5)	50.2(+1.2)	50.6(+1.6)
3	1+BagWM_ED	52.3(+0.9)	50.9(-0.5)	51.5(+0.1)	51.8(+0.4)	52.5(+1.1)	48.5(-0.5)	48.9(-0.1)	48.6(-0.4)	48.7(-0.3)	49.0(+0.0)
4	1+HC_ED	51.3(-0.1)	50.9(-0.5)	48.3(-3.1)	50.8(-0.6)	49.8(-1.6)	44.9(-4.1)	45.8(-3.2)	45.8(-3.2)	48.7(-0.3)	47.3(-1.7)
5	1+BagWC_ED	51.5(+0.1)	50.8(-0.6)	49.5(-1.9)	51.4(+0.0)	50.3(-1.1)	48.3(-0.7)	46.3(-2.7)	44.0(-5.0)	46.6(-2.4)	44.8(-4.2)
6	2+BagWM_ED	54.3(+2.9)	53.2(+1.8)	53.2(+1.8)	54.0(+2.6)	53.8(+2.4)	52.5(+3.5)	51.4(+2.4)	50.6(+1.6)	50.0(+1.0)	48.6(-0.4)
7	6+HC_ED	53.4(+2.0)	52.3(+0.9)	52.7(+1.3)	54.2(+2.8)	53.1(+1.7)	50.5(+1.5)	50.9(+1.9)	48.4(-0.6)	50.0(+1.0)	48.9(-0.1)
8	7+BagWC_ED	53.4(+2.0)	52.2(+0.8)	50.8(-0.6)	53.5(+2.1)	53.6(+2.2)	49.2(+0.2)	50.7(+1.7)	49.2(+0.2)	47.9(-1.1)	49.5(+0.5)

Table 2: In-domain and Out-of-domain performance for different embedding features. The cells in bold are the best results.

to facilitate system comparison later. We evaluate C&W word embeddings with 25, 50 and 100 dimensions as well as HLBL word embeddings with 50 and 100 dimensions that are introduced in Turian et al. (2010) and can be downloaded here[4]. The fact that we utilize the large, general and unbiased resources generated from the previous works for evaluation not only helps to verify the effectiveness of the resources across different tasks and settings but also supports our setting of single-system DA.

We use the ACE 2005 corpus for DA experiments (as in Plank and Moschitti (2013)). It involves 6 relation types and 6 domains: broadcast news (bn), newswire (nw), broadcast conversation (bc), telephone conversation (cts), weblogs (wl) and usenet (un). We follow the standard practices on ACE (Plank and Moschitti, 2013) and use **news (the union of bn and nw)** as the source domain and **bc**, **cts** and **wl** as our target domains. We take half of bc as the only target development set, and use the remaining data and domains for testing purposes (as they are small already). As noted in Plank and Moschitti (2013), the distributions of relations as well as the vocabularies of the domains are quite different.

6.2 Evaluation of Word Embedding Features

We investigate the effectiveness of word embeddings on lexical features by following the procedure described in Section 5.2. We test our system on two scenarios: In-domain: the system is trained and evaluated on the source domain (bn+nw, 5-fold cross validation); Out-of-domain: the system is trained on the source domain and evaluated on the target development set of bc (bc dev). Table 2 presents the F measures of this experiment[5] (the suffix *ED* in lexical group names is to indicate the embedding features).

From the tables, we find that for C&W and HLBL embeddings of 50 and 100 dimensions, the most effective way to introduce word embeddings is to add embeddings to the heads of the two mentions (row 2; both in-domain and out-of-domain) although it is less pronounced for HLBL embedding with 50 dimensions. Interestingly, for C&W embedding with 25 dimensions, adding the embedding to both heads and words of the two mentions (row 6) performs the best for both in-domain and out-of-domain scenarios. This is new compared to the word cluster features where the heads of the two mentions are always the best places for augmentation (Sun et al., 2011). It suggests that a suitable amount of embeddings for words in the mentions might be useful for the augmentation of the heads and inspires further exploration. Introducing embeddings to words of mentions alone has mild impact while it is generally a bad idea to augment chunk heads and words in the contexts.

Comparing C&W and HLBL embeddings is somehow more complicated. For both in-domain and out-of-domain settings with different numbers of dimensions, C&W embedding outperforms HLBL embedding when only the heads of the mentions are augmented while the degree of negative impact of HLBL embedding on chunk heads as well as context words seems less serious than C&W's. Regarding the incremental addition of features (rows 6, 7, 8), C&W is better for the out-of-domain performance when 50 dimensions are used, whereas HLBL (with both 50 and 100 dimensions) is more effective for the in-domain setting. For the next experiments, we will apply the C&W embedding of 50 dimensions to the heads of the mentions for its best out-of-domain performance.

[4] http://metaoptimize.com/projects/wordreprs/

[5] All the in-domain improvement in rows 2, 6, 7 of Table 2 are significant at confidence levels $\geq 95\%$.

6.3 Domain Adaptation with Word Embeddings

This section examines the effectiveness of word representations for RE across domains. We evaluate word cluster and embedding (denoted by ED) features by adding them individually as well as simultaneously into the baseline feature set. For word clusters, we experiment with two possibilities: (i) only using a single prefix length of 10 (as Plank and Moschitti (2013) did) (denoted by WC10) and (ii) applying multiple prefix lengths of 4, 6, 8, 10 together with the full string (denoted by WC). Table 3 presents the system performance (F measures) for both in-domain and out-of-domain settings.

System	In-domain	bc	cts	wl
Baseline(B)	51.4	49.7	41.5	36.6
B+WC10	52.3(+0.9)	50.8(+1.1)	45.7(+4.2)	39.6(+3)
B+WC	53.7(+2.3)	52.8(+3.1)	46.8(+5.3)	41.7(+5.1)
B+ED	54.1(+2.7)	52.4(+2.7)	46.2(+4.7)	42.5(+5.9)
B+WC+ED	**55.5(+4.1)**	**53.8(+4.1)**	**47.4(+5.9)**	**44.7(+8.1)**

Table 3: Domain Adaptation Results with Word Representations. All the improvements over the baseline in Table 3 are significant at confidence level $\geq 95\%$.

The key observations from the table are:

(i): The baseline system achieves a performance of 51.4% within its own domain while the performance on target domains bc, cts, wl drops to 49.7%, 41.5% and 36.6% respectively. Our baseline performance is worse than that of Plank and Moschitti (2013) only on the target domain cts and better in the other cases. This might be explained by the difference between our baseline feature set and the feature set underlying their kernel-based system. However, the performance order across domains of the two baselines are the same. Besides, the baseline performance is improved over all target domains when the system is enriched with word cluster features of the 10 prefix length only (row 2).

(ii): Over all the target domains, the performance of the system augmented with word cluster features of various granularities (row 3) is superior to that when only cluster features for the prefix length 10 are added (row 2). This is significant (at confidence level $\geq 95\%$) for domains bc and wl and verifies our assumption that various granularities for word cluster features are more effective than a single granularity for domain adaptation of RE.

(iii): Row 4 shows that word embedding itself is also very useful for domain adaptation in RE since it improves the baseline system for all the target domains.

(iv): In row 5, we see that the addition of both word cluster and word embedding features improves the system further and results in the best performance over all target domains (this is significant with confidence level $\geq 95\%$ in domains bc and wl). The result suggests that word embeddings seem to capture different information from word clusters and their combination would be effective to generalize relation extractors across domains. However, in domain cts, the improvement that word embeddings provide for word clusters is modest. This is because the RCV1 corpus used to induce the word embeddings (Turian et al., 2010) does not cover spoken language words in cts very well.

(v): Finally, the in-domain performance is also improved consistently demonstrating the robustness of word representations (Plank and Moschitti, 2013).

6.4 Domain Adaptation with Regularization

All the experiments we have conducted so far do not apply regularization for training. In this section, in order to evaluate the effect of regularization on the generalization capacity of relation extractors across domains, we replicate all the experiments in Section 6.3 but apply regularization when relation extractors are trained[6]. Table 4 presents the results.

System	In-domain	bc	cts	wl
Baseline(B)	56.2	55.5	48.7	42.2
B+WC10	57.5(+1.3)	57.3(+1.8)	52.3(+3.6)	45.0(+2.8)
B+WC	58.9(+2.7)	58.4(+2.9)	52.8(+4.1)	47.3(+5.1)
B+ED	58.9(+2.7)	59.5(+4.0)	52.6(+3.9)	48.6(+6.4)
B+WC+ED	**59.4(+3.2)**	**59.8(+4.3)**	**52.9(+4.2)**	**49.7(+7.5)**

Table 4: Domain Adaptation Results with Regularization. All the improvements over the baseline in Table 4 are significant at confidence level $\geq 95\%$.

For this experiment, every statement in (ii), (iii), (iv) and (v) of Section 6.3 also holds. More importantly, the performance in every cell of Table 4 is significantly better than the corresponding cell in Table 3 (5% or better gain in F measure, a significant improvement at confidence level $\geq 95\%$). This demonstrates the effectiveness of regularization for RE in general and for domain adaptation of RE specifically.

[6]We use a L2 regularizer with the regularization parameter of 0.5 for its best experimental results.

References

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2001. *A Neural Probabilistic Language Model*. In Advances in Neural Information Processing Systems (NIPS'13), pages 932-938, MIT Press, 2001.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. *A Neural Probabilistic Language Model*. In Journal of Machine Learning Research (JMLR), 3, pages 1137-1155, 2003.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. *Domain Adaptation with Structural Correspondence Learning*. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. *Biographies, Bollywood, Boom-boxes, and Blenders: Domain Adaptation for Sentiment Classification*. In Proceedings of the ACL, pages 440-447, Prague, Czech Republic, June 2007.

John Blitzer, Dean Foster, and Sham Kakade. 2011. *Domain Adaptation with Coupled Subspaces*. In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, pages 173-181, Fort Lauderdale, FL, USA.

Elizabeth Boschee, Ralph Weischedel, and Alex Zamanian. 2005. *Automatic Information Extraction*. In Proceedings of the International Conference on Intelligence Analysis.

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*. In Journal of Computational Linguistics, Volume 18, Issue 4, pages 467-479, December 1992.

Razvan C. Bunescu and Raymond J. Mooney. 2005a. *A Shortest Path Dependency Kenrel for Relation Extraction*. In Proceedings of HLT/EMNLP.

Razvan C. Bunescu and Raymond J. Mooney. 2005b. *Subsequence Kernels for Relation Extraction*. In Proceedings of NIPS.

Yee S. Chan and Dan Roth. 2010. *Exploiting Background Knowledge for Relation Extraction*. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 152-160, Beijing, China, August.

Ronan Collobert and Jason Weston. 2008. *A Unied Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning*. In International Conference on Machine Learning, ICML, 2008.

Hal Daumé III. 2007. *Frustratingly Easy Domain Adaptation*. In Proceedings of the ACL, pages 256-263, Prague, Czech Republic, June 2007.

Hal Daumé III, Abhishek Kumar and Avishek Saha. 2010. *Co-regularization Based Semi-supervised Domain Adaptation*. In Advances in Neural Information Processing Systems 23 (2010).

Ralph Grishman, David Westbrook and Adam Meyers. 2005. *NYU's English ACE 2005 System Description*. ACE 2005 Evaluation Workshop.

Fei Huang and Alexander Yates. 2010. *Exploring Representation-Learning Approaches to Domain Adaptation*. In Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, pages 23-30, Uppsala, Sweden, July 2010.

Jing Jiang and ChengXiang Zhai. 2007a. *A Systematic Exploration of the Feature Space for Relation Extraction*. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'07), pages 113-120, 2007.

Jing Jiang and ChengXiang Zhai. 2007b. *A Two-stage Approach to Domain Adaptation for Statistical Classifiers*. In Proceedings of the ACM 16th Conference on Information and Knowledge Management (CIKM'07), pages 401-410, 2007.

Nanda Kambhatla. 2004. *Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Information Extraction*. In Proceedings of ACL-04.

Daniel Khashabi. 2013. *On the Recursive Neural Networks for Relation Extraction and Entity Recognition*. Technical Report (May, 2013), UIUC.

Pavel Kuksa, Yanjun Qi, Bing Bai, Ronan Collobert, Jason Weston, Vladimir Pavlovic, and Xia Ning. 2010. *Semi-Supervised Abstraction-Augmented String Kernel for Multi-Level Bio-Relation Extraction*. In Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases, Part II (ECML PKDD'10), pages 128-144, 2010.

Andrew L. Maas and Andrew Y. Ng. 2010. *A Probabilistic Model for Semantic Word Vectors*. In NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2010.

Andriy Mnih and Geoffrey Hinton. 2007. *Three new Graphical Models for Statistical Language Modelling*. In Proceedings of ICML'07, pages 641-648, Corvallis, OR, 2007.

Andriy Mnih and Geoffrey Hinton. 2009. *A Scalable Hierarchical Distributed Language Model*. In NIPS, page 1081-1088.

Truc-Vien T. Nguyen, Alessandro Moschitti, and Giuseppe Riccardi. 2009. *Convolution Kernels on Constituent, Dependency and Sequential Structures for Relation Extraction*. In Proceedings of EMNLP 09, pages 1378-1387, Stroudsburg, PA, USA.

Barbara Plank and Alessandro Moschitti. 2013. *Embedding Semantic Similarity in Tree Kernels for Domain Adaptation of Relation Extraction*. In Proceedings of the ACL 2013, pages 1498-1507, Sofia, Bulgaria.

Longhua Qian, Guodong Zhou, Qiaoming Zhu and Peide Qian. 2008. *Exploiting Constituent Ddependencies for Tree Kernel-based Semantic Relation Extraction*. In Proceedings of COLING, pages 697-704, Manchester.

Tobias Schnabel and Hinrich Schütze. 2014. *FLORS: Fast and Simple Domain Adaptation for Part-of-Speech Tagging*. In Transactions of the Association for Computational Linguistics, 2 (2014), pages 1526.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. *Semantic Compositionality through Recursive Matrix-Vector Spaces*. In Proceedings EMNLP-CoNLL'12, pages 1201-1211, Jeju Island, Korea, July 2012.

Ang Sun, Ralph Grishman, and Satoshi Sekine. 2011. *Semi-supervised Relation Extraction with Large-scale Word Clustering*. In Proceedings of ACL-HLT, pages 521-529, Portland, Oregon, USA.

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 (ACL'10), pages 384-394, Uppsala, Sweden, July, 2010.

Min Xiao and Yuhong Guo. 2013. *Domain Adaptation for Sequence Labeling Tasks with a Probabilistic Language Adaptation Model*. In Proceedings of the 30th International Conference on Machine Learning (ICML-13), pages 293-301, 2013.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. *Kernel Methods for Relation Extraction*. Journal of Machine Learning Research, 3:10831106.

Min Zhang, Jie Zhang, Jian Su, and GuoDong Zhou. 2006. *A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features*. In Proceedings of COLING-ACL-06, pages 825-832, Sydney.

Guodong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. *Exploring various Knowledge in Relation Extraction*. In Proceedings of ACL'05, pages 427-434, Ann Arbor, USA, 2005.