A Corpus to Learn Refer-to-as Relations for Nominals

Wasi Uddin Ahmad and Kai-Wei Chang
Department of Computer Science, University of California, Los Angeles, California, USA
{wasiahmad, kwchang}@cs.ucla.edu

Abstract

Continuous representations for words or phrases, trained on large unlabeled corpora are proved very useful for many natural language processing tasks. While these vector representations capture many fine-grained syntactic and semantic regularities among words or phrases, it often lacks coreferential information which is useful for many downstream tasks like information extraction, text summarization etc. In this paper, we argue that good word and phrase embeddings should contain information for identifying refer-to-as relationship and construct a corpus from Wikipedia to generate coreferential neural embeddings for nominals. The term *nominal* refers to a word or a group of words that functions like a noun phrase. In addition, we use coreference resolution as a proxy to evaluate the learned neural embeddings for noun phrases. To simplify the evaluation procedure, we design a coreferential phrase prediction task where the learned nominal embeddings are used to predict which candidate nominals can be referred to a target nominal. We further describe how to construct an evaluation dataset for such task from well known OntoNotes corpus and demonstrate encouraging baseline results.

Keywords: Nominals, Coreference, Refer-to-as relation

1. Introduction

Understanding relations between words and phrases is a long-standing problem in natural language processing. Various resources are collected and utilized in order to understand different types of relations between words, including synonymy, antonymy (Turney, 2008) and hierarchical relationships such as hyponymy and hypernymy (Fu et al., 2014). Coreference (a.k.a refer-to-as) relation is another important type of relations between words and phrases and has a wide range of potential applications. Previous work (Feng et al., 2015) found both textual and visual information helpful to learn refer-to-as relations between words. However, annotated data for coreferent phrases are missing. In this work, we develop a dataset from Wikipedia which can aid in learning and evaluating refer-to-as relations between a group of words which act as a noun phrase. Furthermore, the developed dataset can be leveraged to construct semantic space representations for the coreferential nominals.

Learning semantic representations for words from large unlabeled corpora (ex., Wikipedia) using co-occurrence statistics has a long history in natural language processing (Deerwester et al., 1990; Lund and Burgess, 1996; Collobert and Weston, 2008). More recent works (Mikolov et al., 2013; Pennington et al., 2014) uses log-bilinear models to learn continuous representations of words on large corpora efficiently. While these vector representations capture fine-grained syntactic and semantic regularities among words or phrases, it often lacks coreferential information. For example, "phd student" and "graduate fellow" can be co-referred to each other and this relationship should be recognized by semantic representations.

Refer-to-as relation information can benefit many natural language processing applications such as question answering (Morton, 1999), information extraction (Humphreys et al., 1997; Zelenko et al., 2004) etc. So, in this paper, we focus on the task of resolving refer-to-as relation between nominals. We design a coreferential phrase prediction task by simplifying the coreference resolution task to evaluate the utility of our proposed corpus. The automatic resolution of identifying surface forms (a.k.a mentions) which co-refer to the same abstract entity is a challenging task with a long history in computational linguistics. For example, given a paragraph, "A female motorist wearing a blue shirt abruptly made a left turn, ignoring the officer's attempt to initiate a traffic stop. The driver continued to drive erratically to Annapolis Road.", an automatic system has to recognize "A female motorist wearing a blue shirt" and "the driver" refer to the same entity. To deal with such a task, the underlying system has to decide whether two noun phrases are compatible and whether they can narratively replace each other. This requires a high-level understanding of the semantics of words and phrases.

In order to learn representations which can capture the coreferential relationship between nominals, we propose a corpus extracted from Wikipedia. To evaluate the learned representation of nominals from our proposed corpus, we use a simplified form of coreference resolution task and use it as a proxy to evaluate the learned representation of noun phrases. In the following, we provide an example to illustrate the advantage of utilizing coreferential information from Wikipedia in resolving nominal coreference.

Example: The September 11 attacks *by al-Qaeda killed 2,996 people and caused at least $10 billion in property and infrastructure damage. It was* the deadliest incident *for firefighters and law enforcement officers in the history of the United States.*

Resolving the nominal mention, *the deadliest incident* to the mention *September 11 attacks* requires help from external knowledge source. For example, if we know *incident* can be linked to *attacks* from a knowledge source, ex., Wikipedia, then we can resolve the coreference in the provided example. We argue that good representations of noun phrases should contain sufficient information to identify such coreferential relationship. So, we train deep neural networks to learn phrase representations based on our pro-

posed corpus and design a coreferential phrase prediction task to evaluate the learned representations. This evaluation is complementary to common word embeddings evaluation on synonymy, hypernymy and hyponymy (Fu et al., 2014) and is more close to human's perception of word and phrase meanings.

The traditional coreference resolution task is a supervised clustering task. Given a document, a system clusters all the mentions in the article into equivalent classes, such that each class contains mentions refer to the same entity. Despite we can plug-in word and/or phrase representations into a coreference resolution system and evaluate the soundness of the representations based on the end performance of the system but it is hard to measure the contribution of the individual word/phrase representations. As our main goal is to evaluate representations of noun phrases, we specifically focus on the nominals since previous works (Durrett and Klein, 2014) show that resolving nominal mentions are one of the hardest categories in coreference resolution.

Therefore, we propose a ranking evaluation procedure based on the intuition that many state-of-the-art approaches are based on mention ranking model (Denis and Baldridge, 2008). Given a target mention and a list of candidate mentions, the goal in our setting is to rank the mentions in the candidate list based on how likely it is co-referred with the target mention without considering the context. In this way, we can directly compare the contributions of the mention representations in a model. We propose a phrase level task where we consider the entire mention boundaries [1] as a noun phrase. We demonstrate how to construct such mention ranking evaluation data from Ontonotes v5.0 corpus and present the results of preliminary experiments as a proof-of-concept.

2. Dataset Construction

We construct our proposed corpus on the Wikipedia dump from March 06, 2016 which contains 16.4M (approx.) articles where 6.5M (approx.) articles are redirected to another Wikipedia article. We considered rest of the 9.9M non-redirected Wikipedia articles to generate the dataset that will be used to learn the coreferential relationship between nominals. We treat each article on Wikipedia as representing an entity (or concept or idea), and the anchor text of inlinks as a mention of the entity. All the wiki links present in a Wikipedia article are extracted and tagged with appropriate part-of-speech using Stanford log-linear POS tagger. We consider the anchor text of the hyperlinks as mentions. Since we are focusing on nominals, we tag the noun phrases to identify nominals from rest of the noun phrases. Nominals are derived from the noun mentions by following a simple rule, *all non-capitalized nouns are nominals*. For example, *daughter* is counted as a nominal but *Professor* is not. The development tool along with the constructed corpus is publicly available. [2]

Number of articles	16,388,870
Number of redirected articles	6,466,828
Number of non-redirected articles	9,922,042
Unique noun mentions	26,660,798
Unique nominal mentions	2,512,347
Unique nominal mentions ($1 \leq$ mention length ≤ 30)	1,428,441

Table 1: Corpus description extracted from Wikipedia

Additionally, we consider the nominals which do not contain noun phrase but are linked to a Wikipedia article whose title contains noun phrase. For example, *self-governed* mention is linked to an article titled as *self-governance* is counted though *self-governed* is an adjective but *self-governance* is a noun phrase. Finally, nominal mentions with length 1 or more than 30 are filtered out for our experiment. Table 1 lists the complete details of the generated dataset.

3. Learning Phrase Embeddings

We propose to use coreference relationship to evaluate word and phrase embedding. It is important to note that, co-occurrence and coreference are not the same concepts. For example, "lazy dog" and "attractive cat" are very close in the low dimensional embedding space because of high co-occurrence of the words "dog" and "cat" but they cannot be co-referred. On the other hand, "phd candidate" and "graduate student" can be co-referred to each other.

As we mentioned, we reduce the coreference resolution problem to an antecedent ranking problem, since many state-of-the-art models (Wiseman et al., 2015; Chang et al., 2013; Durrett and Klein, 2013) are a variant of the mention-ranking model (Denis and Baldridge, 2008). We define the antecedent ranking problem as, given a target mention, the goal is to rank all antecedent mentions in the same document based on predicted scores such that the antecedent mentions referred to the target mention are ranked on the top. We propose to evaluate phrase embeddings to estimate the usefulness of embeddings to train supervised learning algorithms for antecedent ranking. We train an end-to-end model to produce phrase embeddings based on the training set, tune the model on the development set, and compute accuracy of the ranking based on evaluation dataset. We allow a model to use a pre-trained word embedding for initialization purpose. We construct the training and development dataset from the corpus we generated from Wikipedia and the evaluation dataset based on OntoNotes V5.0 corpus (Hovy et al., 2006) that used in the CoNLL shared task 2012 for coreference resolution (Pradhan et al., 2012). Table 2 lists the complete details of the extracted dataset.

We get coreferential mention clusters from the corpus we generate as described in section 2. Negative examples are

[1] Different corpus may have different definitions of mention boundaries. For example, Ontonotes defines the largest noun phrase that represents an entity as mention, while ACE annotation uses the shortest noun phrase to identify a mention. In this work, we follow the definition in Ontonotes corpus.

[2] https://github.com/wasiahmad/mining_wikipedia/tree/master/WikiMiner

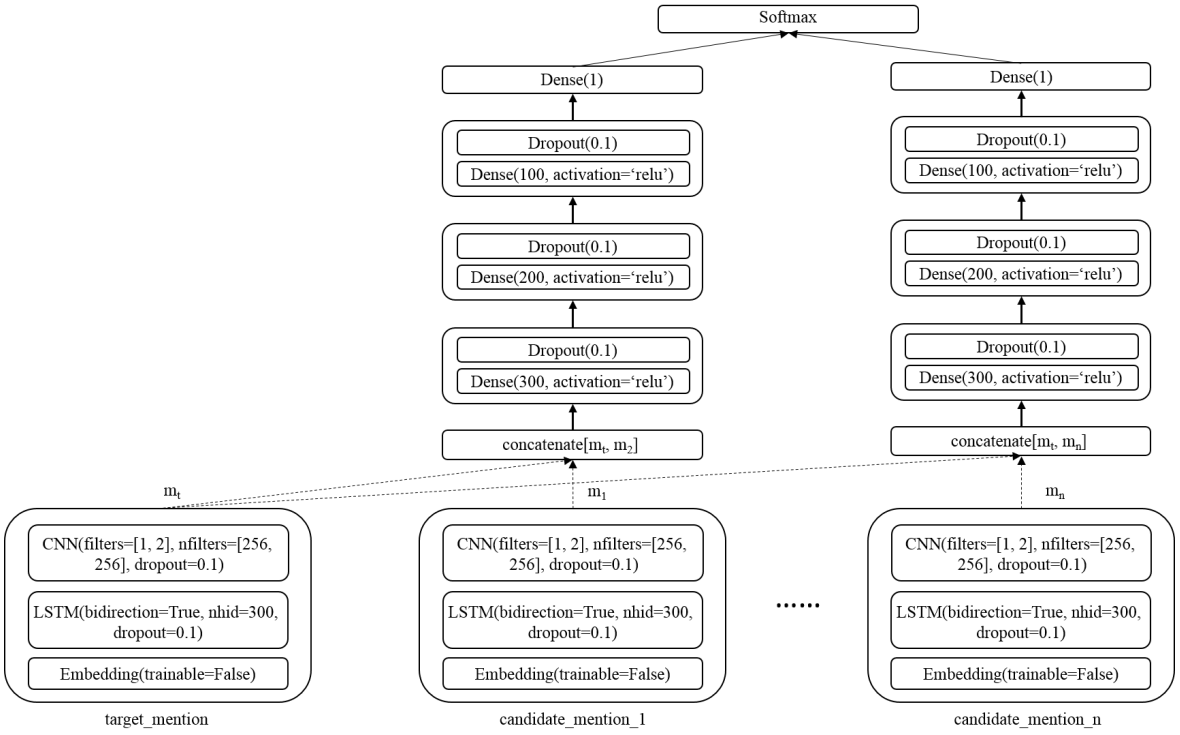

Figure 1: Neural network architecture to rank candidate mentions of a target mention

Train (src: Wikipedia)	Total nominal coref. chain	78,665
	Avg. candidates per chain	24
	Total unique terms	35,939
Development (src: Wikipedia)	Total nominal coref. chain	8,354
	Avg. candidates per chain	18
	Total unique terms	6,686
Test (src: CoNLL)	Total nominal coref. chain	623
	Avg. candidates per chain	12
	Total unique terms	2,839

Table 2: Data Description

selected using random sampling based on cosine similarity distribution. To compute similarity, we simply use pre-trained word embeddings (Pennington et al., 2014). Examples of train/development dataset is provided in table 3.

Baselines: The "Mention Embeddings" (an unsupervised approach) baseline simply take the average of the word vectors in a phrase as the phrase embedding, and compute the cosine similarity to score phrase pairs. Formally, given the embeddings of n_p words of a phrase p, $w_1, \ldots, w_{n_p}$, the phrase embedding $E(p)$ is:

$$E(p) = \frac{1}{n_p} \sum_{k=1}^{n_p} w_k$$

where $E(p)$, $w_k \in R^{d_e}$ and d_e is a hyper-parameter, indicating word embedding size. Then, we compute the similarity score of the phrase pair as follows.

$$\text{Sim}(p_1, p_2) = \text{cosine}(p_1, p_2) = \frac{p_1^T p_2}{||p_1||||p_2||}$$

The "Mention Embeddings + FFNN" baseline construct mention representations like previous baseline but compute score using a two-layer feed-forward neural network.

$$\text{Sim}(p_1, p_2) = \sigma(u^T \tanh(W[E(p_1), E(p_2)] + b))$$

where $W \in R^{d_e \times d_e}, b, u \in R^{d_e}$, and $[E(p_1), E(p_2)]$ represents concatenation of the phrase embedding pair.

Inspired by (Chiu and Nichols, 2016; Ma and Hovy, 2016), we also provide a more sophisticated baseline by using bidirectional long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and convolutional neural network (CNN) (LeCun et al., 1998; Kalchbrenner et al., 2014) in combinations to construct mention representation and using a feed-forward neural network to score mention antecedent pairs. A general architecture for the proposed baseline methods is shown in Fig. 1.

We use a shallow bi-directional LSTM with hidden size $\tilde{h}$ to encode contextual embeddings $\tilde{p}_t$ of each word in the phrase,

$$\overrightarrow{h}_t = LSTM(\overrightarrow{h}_{t-1}, w_t), t = 1, \ldots, n_p$$
$$\overleftarrow{h}_t = LSTM(\overleftarrow{h}_{t+1}, w_t), t = n_p, \ldots, 1$$
$$\tilde{p}_t = [\overrightarrow{h}_t, \overleftarrow{h}_t]$$

Where $\tilde{p}_t \in R^h$ and $h = 2\tilde{h}$.

To construct mention embeddings, we apply convolution operation on the contextual embeddings, $\tilde{p}_t$. We can view the convolution operation as sliding window based feature extractor which captures the n-gram contextual features. A convolution operation involves a filter $w \in R^N (N = nd_e)$, which is applied to a window of n words to produce an n-gram contextual feature. The filter is applied to each possible window of n words in the mention to produce a feature

Target Mention	Positive Candidates	Negative Candidates
protein sequence	amino acid sequencing, chain of amino acids, peptide sequence, protein primary structure	metabolic enzymes, biological mutations, periodic sequence, nucleotide sequence
general election	whole coalition, upcoming election, the previous election, election campaign, legislative election	the constitutional amendment, election win, the presidential election, democratic political values
aerial bomb	aerial bombardment, bombing, bomb attack	nuclear bomb technology, terror attacks, attack ground targets, atomic weapon
highway construction	roads, road building equipment, road work construction, street construction, road building	highway marker, construction yard, railway and highway bridge, construction superintendent

Table 3: Example of positive and negative coreference clusters generated from Wikipedia

Model	NLL-Loss	MAP	P@1	P@5	R@1	R@5
Mention Embeddings	1.7389	**0.5452**	**0.5185**	0.2374	**0.3715**	0.7630
Mention Embeddings + FFNN	1.7836	0.4632	0.4995	0.2317	0.3516	0.7888
Bidirectional-LSTM + CNN + FFNN	**1.6731**	0.4884	0.4719	**0.2475**	0.3476	**0.8025**

Table 4: Performance of baseline methods.

map. Then max pooling is applied over the feature map to select one feature for one filter. In this way, all the contextual features generated by a predefined number of filters are concatenated to produce the mention representation.

A two-layer feed-forward neural network is used to score mention antecedent pairs. Embedding of the mention and it's candidate antecedent are concatenated and given as an input to the feed-forward network. To add non-linearity after each layer, we use rectified linear units in the feed-forward network.

Hyper-parameter Tuning. We carefully tune parameters on the development set. Parameters of the model were learned using mini-batch SGD with Adam (Kingma and Ba, 2014) for optimization with the two momentum parameters set to 0.9 and 0.999 respectively. Initial learning rate was set to 0.001. We use mini-batches of size 64 and early stopping criterion to stop training if validation accuracy does not improve for 3 iterations. For the hidden size of BiLSTM, we consider the range $[150, 300, 600]$ and found 300 results in best performance. We use convolution filters of size 1 and 2 with 256 feature maps each. Gradient clipping technique (5.0) (Graves, 2013) and dropout (0.1) (Srivastava et al., 2014) were used. We use pre-trained GloVe embeddings (Pennington et al., 2014) and fix them during training. Out-of-vocabulary words were initialized with zero vectors.

4. Evaluation Metrics and Baseline Results

In this section, we present the details of evaluation metrics and the performances of baseline methods.

Evaluation Metrics. In tradition, researchers treat coreference resolution as a supervised clustering problem and evaluate system performance by clustering metrics (Pradhan et al., 2014). However, this evaluation metric does not align with our goal of ranking antecedents for a given target mention. Therefore, we evaluate the performance by Mean Average Precision (MAP), Precision at k ($P@k$), Recall at k ($R@k$). We also report the negative log-likelihood loss for the baseline methods.

Results. The performance comparison based on the evaluation dataset between baseline models is presented in Table 4. For our proposed phrase embeddings evaluation task, averaging word vectors is a strong baseline, even without accessing training dataset, it achieves better results to the trained neural network based approaches. This could be due to the noise coming from the unlabeled corpus generated from Wikipedia. We need to deal with the noise to capture a better coreferential relationship between phrases and we are leaving this as our future work.

5. Related Works

Distributed word representations, also known as, word embeddings, typically represent words with dense, low-dimensional and real-valued vectors. Word embeddings have been empirically shown to preserve linguistic information, such as the semantic relationship between words (Mikolov et al., 2013; Pennington et al., 2014) and it helps to learn algorithms to perceive underlying semantics of the targeted task. Over the past few years, researchers have been studying different ways for evaluating word embeddings, including using hypernym-hyponym relation (Fu et al., 2014), word similarity task (Levy and Goldberg, 2014), word analogy taks (Levy et al., 2015), POS tagging task (Lin et al., 2015) and phrase-based machine translation (Zou et al., 2013). Our proposed evaluation approach is complementary to the previous ones. Besides, it is suitable to be used for evaluating phrase embedding.

Our work is inspired by previous works on supervised coreference research which show that incorporating external knowledge can improve the performance of a coreference system (CR). A variety of approaches (Ng, 2007; Ponzetto and Strube, 2006; Haghighi and Klein, 2009) have

been shown to benefit from using external resources such as Wikipedia (Strube and Ponzetto, 2006; Ponzetto and Strube, 2007; Singh et al., 2012; Spitkovsky and Chang, 2012), WordNet (Harabagiu et al., 2001; Soon et al., 2001) or YAGO (Suchanek et al., 2007). (Rahman and Ng, 2011) examined the utility of three major sources of world knowledge and applied them to two learning-based coreference models and found improved performance when knowledge extracted from different sources are exploited in combination rather than individually. Also, previous works (Ogrodniczuk, 2013) verified that nominal facts extracted from world knowledge resources effectively perform and can be used as a source of pragmatic knowledge for coreference resolution.

6. Conclusion and Future Work

In this work, we propose a corpus to learn refer-to-as relations for nominals extracted from Wikipedia in an unsupervised way. Also, we carry out an extrinsic evaluation of phrase embeddings which can aid in resolving nominal coreference. We simplified the coreference resolution problem and presented it as an antecedent ranking task. We have provided several baseline techniques to demonstrate the efficacy of phrase embeddings in resolving nominal mentions. In future, we are interested to investigate, whether word and phrase embeddings can be trained in such a way that learned representations can capture coreferential relationship along with other linguistics regularities and patterns.

Acknowledgements

This work was supported in part by National Science Foundation Grant IIS-1760523 and an NVIDIA Hardware Grant.

7. Bibliographical References

Chang, K.-W., Samdani, R., and Roth, D. (2013). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Chiu, J. and Nichols, E. (2016). Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Denis, P. and Baldridge, J. (2008). Specialized models and ranking for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 660–669. Association for Computational Linguistics.

Durrett, G. and Klein, D. (2013). Easy victories and uphill battles in coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, October. Association for Computational Linguistics.

Durrett, G. and Klein, D. (2014). A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Feng, S., Ravi, S., Kumar, R., Kuznetsova, P., Liu, W., Berg, A. C., Berg, T. L., and Choi, Y. (2015). Refer-to-as relations as semantic knowledge. In *AAAI*, pages 2160–2166.

Fu, R., Guo, J., Qin, B., Che, W., Wang, H., and Liu, T. (2014). Learning semantic hierarchies via word embeddings. In *ACL (1)*, pages 1199–1209.

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Haghighi, A. and Klein, D. (2009). Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1152–1161. Association for Computational Linguistics.

Harabagiu, S. M., Bunescu, R. C., and Maiorano, S. J. (2001). Text and knowledge mining for coreference resolution. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Humphreys, K., Gaizauskas, R., and Azzam, S. (1997). Event coreference for information extraction. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 75–81. Association for Computational Linguistics.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Lin, C.-C., Ammar, W., Dyer, C., and Levin, L. (2015). Unsupervised pos induction with word embeddings. *arXiv preprint arXiv:1503.06760*.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

Ma, X. and Hovy, E. H. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. The Association for Computer Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Morton, T. S. (1999). Using coreference for question answering. In *Proceedings of the Workshop on Coreference and its Applications*, pages 85–89. Association for Computational Linguistics.

Ng, V. (2007). Shallow semantics for coreference resolution. In *IJcAI*, volume 2007, pages 1689–1694.

Ogrodniczuk, M. (2013). Discovery of common nominal facts for coreference resolution: Proof of concept. In *Mining Intelligence and Knowledge Exploration*, pages 709–716. Springer.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ponzetto, S. P. and Strube, M. (2006). Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 192–199. Association for Computational Linguistics.

Ponzetto, S. P. and Strube, M. (2007). Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Intell. Res.(JAIR)*, 30:181–212.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

Pradhan, S., Luo, X., Recasens, M., Hovy, E., Ng, V., and Strube, M. (2014). Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, June.

Rahman, A. and Ng, V. (2011). Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 814–824. Association for Computational Linguistics.

Singh, S., Subramanya, A., Pereira, F., and McCallum, A. (2012). Wikilinks: A large-scale cross-document coreference corpus labeled via links to wikipedia. *University of Massachusetts, Amherst, Tech. Rep. UM-CS-2012-015*.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Spitkovsky, V. I. and Chang, A. X. (2012). A cross-lingual dictionary for english wikipedia concepts. In *LREC*, pages 3168–3175.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Strube, M. and Ponzetto, S. P. (2006). Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pages 1419–1424.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Turney, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 905–912. Association for Computational Linguistics.

Wiseman, S. J., Rush, A. M., Shieber, S. M., and Weston, J. (2015). Learning anaphoricity and antecedent ranking features for coreference resolution. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics.

Zelenko, D., Aone, C., and Tibbetts, J. (2004). Coreference resolution for information extraction. In *Proceedings of the ACL Workshop on Reference Resolution and its Applications*, pages 9–16.

Zou, W. Y., Socher, R., Cer, D. M., and Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.