TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 03/15 pp586-594
Volume 22, Number 6, December 2017

How Do Pronouns Affect Word Embedding

Tonglee Chung, Bin Xu, Yongbin Liu*, Juanzi Li, and Chunping Ouyang

Abstract: Word embedding has drawn a lot of attention due to its usefulness in many NLP tasks. So far a handful of neural-network based word embedding algorithms have been proposed without considering the effects of pronouns in the training corpus. In this paper, we propose using co-reference resolution to improve the word embedding by extracting better context. We evaluate four word embeddings with considerations of co-reference resolution and compare the quality of word embedding on the task of word analogy and word similarity on multiple data sets. Experiments show that by using co-reference resolution, the word embedding performance in the word analogy task can be improved by around 1.88%. We find that the words that are names of countries are affected the most, which is as expected.

Key words: word embedding; co-reference resolution; representation learning

1 Introduction

Word embedding or distributed word representation can be referred to as a mathematical object associated with a word, usually a vector. It can capture semantic meanings of words from the unlabeled corpus of text and is becoming a vital part of many NLP tasks. Many models have been proposed to learn word representation based on the distributional hypothesis of Harris[1]. This hypothesis can be interpreted as: words that occur in the same contexts tend to have similar meaning[2]. Many word representation methods based on language models have been proposed. Although these studies have shown intriguing results, they do not focus on the names of people, countries, etc. In other words, they only care about the general words and those that appear in evaluation datasets. Although pronouns have been proven to be able to find *male-female* relations using *he-she*. But the names are also very important in some fields, such as news event mining and entity linking, and should be taken seriously. The problem with a person's name is that it is replaced with a pronoun in many parts of text articles so when extracting the word context pair for training the word embedding, some information is lost. Therefore, this paper focuses on the effects that pronouns have on word embedding quality, applies co-reference resolution to word embedding algorithms, and evaluates the word embedding in different data sets in word similarity and word analogy tasks.

To understand how pronouns affect word embedding, the model construction and context selection have to first be understood and clarified. Mikolov et al.[3–7] introduced the CBOW and Skip-gram, two efficient models that learn high quality vector representations of words from large unstructured text corpuses. This work is implemented in the open-source word2vec software. There is also a derivation of this work by Levy and Goldberg[8] that proposed an alternative method to the original linear bag-of-words approach by incorporating syntactic dependency relations. GloVe[9] represents another line of word embedding model called the count-based model[10]. In this paper, we generalize the process that these models extract context for word embedding learning. Then, co-reference resolution is

• Tonglee Chung, Bin Xu, and Juanzi Li are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: tongleechung86@gmail.com; xubin@tsinghua.edu.cn; juanzili@tsinghua.edu.cn.
• Yongbin Liu and Chunping Ouyang are with the School of Computer Science and Technology, University of South China, Hengyang 421001, China. E-mail: qingbinliu@163.com; ouyangcp@126.com.
∗ To whom correspondence should be addressed.
Manuscript received: 2016-12-31; revised: 2017-03-29; accepted: 2017-05-25

included into these four different models to observe the extent that pronouns affect word embedding quality. The quality of word embedding will be evaluated in two separate ways: first, we evaluate our model in the task of word analogy using the linear substructure of word embedding; second, we use multiple data sets introduced by Faruque and Dyer's work[11] for Spearman's ranking.

Pronouns are commonly used in article writing and are targeted for human readers. Human minds have a very comprehensive understanding of context and easily understand its original reference, but pronouns pose a big problem for word embedding computation. So, we propose using co-reference resolution to enhance the quality when inducing word embedding. Pronouns are used very often in articles and writing, and embedding a pronoun is not useful in some NLP tasks because it can refer to almost everything. Pronouns not only provide little help in some tasks, but they may actually add noise to word embedding learning. For example, if we look at the sample sentence from the Wikipedia about Obama, "Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree." If we set the context window to 5 and consider stop words, we will only be able to get the context set of *graduate, Comlumbia, University, Harvard*, and *Law* for the word *Obama*. For more accurate word embedding, the context of *Obama* should also include *community, organizer, Chicago*, etc. But because of the existence of pronouns, these words will not be considered as the context of *Obama*. Some may argue that the word embedding is trained using the large text corpus and can compensate for noise, but statistics show that pronouns are used very often in articles. Our experiments show that replacing pronouns with the original word yields better results in some tasks.

The main contributions of this paper can be summarized as follows:

- We analyze pronouns in the text corpus and their effect when extracting context for word embedding inducing.
- We present a novel word embedding model that takes advantage of co-reference resolved context to improve the word embedding.
- We perform the novel word embedding on two different tasks, word similarity and word analogy. Experiments show that the model is effective for two tasks.

The rest of the paper is organized as follows. We analyze two word embedding context extraction methods in Section 2. Then, Section 3 describes the process of incorporating co-reference resolution to existing methods, followed by experiments and results in Section 4. We discuss the related work in Section 5. Finally, our paper is concluded in Section 6.

2 Word Context Selection

In this section, we look at two diverse ways that word embedding chooses context from the text corpus. The conventional way of generating context is by using a window to choose k-context from the target word, another alternative is to generate the context based on syntactic distance.

2.1 Linear bag-of-words context

A spontaneous way of looking at the distributional hypothesis of Harris is that the context of a target word is closer in physical distance in the sentence. An uncomplicated way of the extracting context is by using a window to select the context that surrounds the target word. This strategy is used by the skip-gram, CBOW, and GloVe embedding models. Special treatment is given to words on either side of the target word. Words with either too high or low frequency are usually removed. In the sample sentence, "Obama is a graduate of Columbia University and Harvard Law School," the contexts of *Obama* using a window of 2 are *graduate* and *Columbia*. A larger window size can capture broad topical content, on the contrary, whereas a smaller window size acquires more concentrated information around the target word.

2.2 Dependency context

Dependency-based context is another way of interpreting the distributional hypothesis. The contexts are extracted based on their syntactic distance. A syntactic dependency parser can resolve syntactically the structure of words in a sentence. The dependency parsed contents are triplets, which include relation label, governor, and dependent. An example of a parsed result of the sample sentence is "*Obama is a graduate of Columbia University and Harvard Law School* is *nsubj (graduate, Obama)*." The context of the target word contains two parts, the dependent part and its relation label. The context of *graduate* is *Obama/nsubj* and the context of *Obama* is $graduate/nsubj^{-1}$ while

−1 indicates that it is an inverse-relation. The word context can be extracted from a word in this fashion and these contexts are syntactically closer to the target word.

3 Word Embedding with the Co-reference Resolved Context

In this part, we propose using the co-reference resolved context for word embedding training and detail our strategy to do so.

3.1 Motivation for co-reference resolution

The motivation behind this work is that we find that pronouns are used very often in articles, especially in the articles about people. People's names are very important in news mining and entity linking, but they are often replaced in the same paragraph whenever they do not cause confusion to human readers. Table 1 shows the number of occurrence for several types of pronouns using the Stanford parser. The total number of pronouns in the whole Wikipedia is 6.7×10^7 while a word count for the whole corpus is around 0.6 billion after per-processing. Even though this is a rather small ratio, the contents that these pronouns replace are mostly names and places. For example, in the Wikipedia article of Obama, the word *Obama* appeared around 200 times while *Obama* was replaced by the pronoun over 200 times. That means that half of the contexts of *Obama* cannot be captured during the context extraction process.

3.2 Embedding with the co-reference resolved context

To address the problem with pronouns, we propose using the co-reference resolved context. The task of co-reference resolution is to derive the correct interpretation of the text for the right individual. Take as an example, the resolution of the sentence, "Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree." A co-reference resolution system[12] can correctly point out that the two occurrences of *he* in sentences 1 and 2 refer to *Obama* and *his* in sentence 2 also refers to *Obama*, this is outputted as a co-reference chain. So, when extracting context for the word with a window of size 2, *served, president, community, organizer, law,* and *degree* (stop words and word with high frequency are ignored.) are the contexts of *Obama*. Including the co-reference resolved context can relax the effect provided when using pronouns in multiple sentences. In this paper, we include the co-reference context into four existing embedding models, namely, skip-gram, CBOW, dependency-based embedding, and GloVe.

Table 1 Estimation of occurrences of distinct types of pronouns in Wikipedia articles.

Pronoun type	Number of occurrences
Personal pronoun	1.1×10^7
Prossessive pronoun	1.2×10^7
wh-determiner	1.7×10^7
wh-pronoun	2.1×10^4
Possessive wh-pronoun	6×10^6
Total	6.7×10^7

3.2.1 Co-reference context with linear form

We demonstrate including the co-reference resolved context to a linear context model using the skip-gram model as an example. The original skip-gram model attempts to estimate the context given a word, while the co-reference model estimates the resolved context instead.

The formal training definition is as follows: D is words vocabulary, C is the contexts vocabulary, $\text{Context}_{\text{coref}}(w)$ represents a non-linear co-reference resolved the context for word w. Our goal is to extract a more accurate context set $\text{Context}_{\text{coref}}(w)$ for word embedding training. The word embedding is represented using a numerical vector $\boldsymbol{v}(w)\in\mathbf{R}^m$, where $\boldsymbol{v}(w)$ is the vector of w, and m is the vector space. The training objective is to maximize the probability that $\text{Context}_{\text{conf}}(w)$ is derived from w:

$$\prod_{w\in D} p(w|\text{Context}_{\text{coref}}(w)) \tag{1}$$

We can simplify the problem by taking the log probability:

$$\sum_{w\in D} \log p(w|\text{Context}_{\text{coref}}(w)) \tag{2}$$

The linear bag-of-words context is used in many word embedding exercises. The context of the word is produced by selecting a $2k$ context using a window of size k around the target word. An example is the sentence, "Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree." When using a window with size 2, the generated contexts are *graduate* and *Columbia*, while the obvious contexts such as *served, president, law*, and *degree*

can not be extracted. We can see that by using pronouns, many valuable contexts are ignored because the contexts are close to the pronoun that refers to *Obama*. The co-reference resolved contexts can solve these problems by replacing the pronoun with the original word.

The process of generating the co-reference resolved context is as follows: use the Stanford CoreNLP System (http://nlp.stanford.edu/software/corenlp.shtml) to generate a list of co-reference chains and parts of the speech for each word. For each co-reference chain, replace a word that has a speech part with any of the following: *personal pronoun, possessive pronoun, wh-determiner, wh-pronoun*, and *possessive wh-pronoun* (part of speech are given in Penn Treebank.) with the representative word (word that represents the co-reference chain.) of the chain. After replacing all pronouns, use the strategy of the linear bag-of-words context to extract the resolved context for each word. Notice that by replacing the context, the word that pronouns referred to can have richer contexts. Other models with linear context extraction methods can be changed to co-reference models using this method.

3.2.2 Co-reference context with dependency based form

Like the linear context, we also formalized the method to include the co-reference context to a dependency model. The formal definition of the training problem can be described as: D is words vocabulary, C is the contexts vocabulary, and $\text{Context}_{\text{dep+coref}}(w)$ represents a dependency-based non-linear word context. Specifically, the context $\text{Context}_{\text{dep+conf}}(w)$ connects to the current word w through dependency on the co-reference word. Our goal is to induce a more accurate context set ($\text{Context}_{\text{dep+coref}}(w)$) to train the word embedding. The word embedding is represented using a numerical vector $\boldsymbol{v}(w) \in \mathbf{R}^m$, where $\boldsymbol{v}(w)$ is the vector of w, and m is the vector space. The training objective is to maximize the probability that $\text{Context}_{\text{dep+conf}}(w)$ is derived from w:

$$\prod_{w \in D} p(w|\text{Context}_{\text{dep+coref}}(w)) \tag{3}$$

We can also simplify the problem by taking the log probability:

$$\sum_{w \in D} \log p(w|\text{Context}_{\text{dep+coref}}(w)) \tag{4}$$

The dependency-based context[8] is an alternative to the linear bag-of-word context, that derives the word context based on the dependency parsed context. Compared to the linear bag-of-word context, the dependency-based contexts choose the contexts based on syntactic distance, which are more inclusive and focused. The dependency-based contexts choose the context in the following format: (m_i, lbl_i) or (m_i, lbl_i^{-1}), where m_i is one of the modifiers of the word, lbl is the dependency relation type of the word and the modifier, and -1 indicates that it is an inverse-relation. Although dependency-based context selects the word contexts that are syntactically closer, it still suffers from pronoun noise that is the same as that of the linear bag-of-words context: original words are replaced by pronouns causing the context to be a pronoun, which produces noise in the training.

The process of adding the co-reference resolved context is as follows: use the Stanford CoreNLP System to generate a list of co-reference chains and the parts of speech for each word. Parse each sentence and derive the word context as $(m_i, \text{lbl}_\text{i})$ or $(m_i, \text{lbl}_\text{i}^{-1})$, where m_i is the modifier of the word, *lbl* is the dependency relation type of the word and modifier, and -1 indicates that it is an inverse-relation. After extracting the dependency context, for each co-reference chain, replace with a word that has parts of speech with the following: *personal pronoun, possessive pronoun, wh-determiner, wh-pronoun,* and *possessive wh-pronoun,* with the representative word of the chain and a coref label and coref^{-1} for the inverse. So, instead of having a context like he/nsubj, we will have it replaced with the context Obama/coref.

During the co-reference process, we observe that co-reference systems produce some strange results: for example, the representation of the two generated co-reference chains is *his* instead of *Mitt Romney* and *his parents* instead of *Mitt Romney and Ann Davis* for the sentence, "Raised in Bloomfield Hills, Michigan, by his parents George and Lenore Romney, Mitt Romney spent two and a half years in France as a Mormon missionary starting in 1966. He married Ann Davies in 1969, and their five children are all sons." We ignore these errors and proceed with the process above and treat it as an error propagated from the co-reference resolver. The same strategy applies to the co-reference context as with the linear form.

4 Experiments and Results

In this section, we evaluate four different embedding models, namely skip-gram, CBOW, dependency-based

embedding (DEP embedding), and GloVe with their co-reference resolved context counterparts. The skip-gram, CBOW, and GloVe use the linear bag-of-words context, while the DEP embedding uses the dependency-based context. Evaluation of the embedding is performed on two different tasks, word similarity and word analogy. Faruqui and Dyer[11] proposed using 12 different datasets to evaluate correlations of word embedding. These datasets include: WS-353, WS-SIM, WS-REL, RG-65, MC-30, MTurk-287, MTurk-771, MEN, YP-130, and Rare-Word. Similarity between a given pair of words is calculated using the *cosine* similarity of two corresponding vectors. Finally, Spearman's rank correlation coefficient is used to rank the similarity of two embedding models. We use these 12 datasets for the word similarity evaluation. Another task is word analogy, also known as a syntactic and semantic relation evaluation where we are given two pairs of tuples of words relations that share a common relation. For example, the analogy of "king is to queen as man is to woman" follows the male-female relation. We encode this in the vector space by the vector equation *king–queen* = *man–woman*.

4.1 Environment setup

When training embedding models using the linear bag-of-words context, we set the window size to 5; however the DEP embedding does not use a window size in its setting, so every word has two contexts. The vector size of all embedding models is set to 50. The minimum word count is set to 50 for all embedding models. In the word2vec setting, we set the number of negative samples to 5. We use Wikipedia's corpus as our training corpus. Before training, we use the Stanford CoreNLP toolkit to parse each sentence, extract the dependency tree, and resolve the documents. All these are based on the tools, so there is no human judgment to favor any of the embedding models. We ignore all errors that are generated by the Stanford CoreNLP toolkit and allow them to propagate to the next step. Finally, the context is extracted by the method described in Section 3 and is used for the training. We use the code provided by word2vec to train the CBOW and skipgram embedding models. The GloVe embedding model is trained using the opensource code provided and the dependency embedding model also by code provided on-line. All models are compared with the co-reference resolved context and original context using Spearman's correlation. Only the skip-gram and CBOW embedding models and their co-reference versions are used in the word analogy evaluation. We also take a closer look at the semantic and syntactic accuracy in the CBOW model by changing the size of the training data. Pre-processing of the data includes some basic steps. In all our models, we use the same step, which includes removal of punctuations and replacing all upper-case letters with lower case letters. We only use paragraphs in the text corpus for our experiment and discard all titles and headings. The co-resolution is performed by paragraph using the Stanford co-reference resolver. In the experiments of variable training data size, we uniformly choose paragraphs with dropout percentages of 90%, 75%, 50%, 25%, and 10% to create data sets of different size.

The purpose of our paper is to evaluate the difference between the original context model and the co-reference resolved context model, not to compete with state-of-art embedding technology. We understand that efficiency has a high priority in embedding training and that co-reference resolution is time consuming. But, with faster computers and paralleled computing, we believe that there is room for co-reference resolved embedding.

We would like the reader to note that during the co-reference resolution, many articles could not be resolved with the co-reference system. Our strategy was to drop these articles, thus resulting in a smaller number of articles when training. So, it is normal that our training set is smaller than the latest Wikipedia dump, but the number of articles is the same for all embedding training. Altogether, our whole dataset contains over 4.6 million documents.

4.2 Word analogy

In this section, we detail the task of word analogy and present our findings.

4.2.1 Task of word analogy

The task of word analogy can be traced back to the work of Mikolov et al.[13] This task utilizes the linear sub-structure of the embedding model to find underlying relations between word pairs. Some examples of these relations include country-capital relation and country-currency relation. For example, *Beijing* is similar to *China* in the same sense that *Paris* is similar to *France*. This is the word analogy task, also known as the question-word task. Given three words, guess the fourth word with the following algebraic operation $X = vector(\text{“}Beijing\text{”}) - vector(\text{“}China\text{”}) + vector(\text{“}Paris\text{”})$. The fourth word should be the word with embedding

closest to X in the cosine distance. Altogether, there are over 7.0×10^4 word pair questions in 14 different categories.

4.2.2 Results for word analogy

The word embeddings for the skip-gram and CBOW models are induced using the above settings, skip-gram_coref and CBOW_coref use the co-reference resolved context and are trained using the same setting. Table 2 shows the results for the word analogy experiment. We can see that in this task, the co-reference context improves the accuracy of semantic word guessing by an average of 1.88%. The improvement is more obvious in name related categories like countries, capitals, and currencies. The reason for this is obvious, many of these names replaced pronouns, and once these pronouns are replaced by the original word, the names will have more abundant and accurate context, thus resulting in higher accuracy. On the other hand, we can see that syntactic accuracy does not improve. This result can be easily predicted because our aim was to provide better embedding for names, and only semantic tasks like countries, capitals and cities have names. We can also see that syntactic task like the nationality adjective that is somewhat related to the names has increased accuracy.

Table 3 shows the semantic accuracy and syntactic accuracy for the CBOW model and its co-reference resolved context embedding model as the training data size increases. The co-reference resolved context embedding model has higher semantic accuracy but a lower syntactic accuracy. Even as the percentage of training data increases, the embedding with the co-reference context shows better results for the semantic task. But we can still see that the syntactic accuracy does not improve. Hence, we can see there is a room to use both the original context and the co-reference context simultaneously to provide better word embedding.

Table 2 Word analogy accuracy. (%)

	Skip-gram	Skip-gram+ coref	CBOW	CBOW+ coref
Capital common countries	68.18	**68.77**	67.59	67.76
Capital world	65.14	63.28	67.11	**69.10**
Currency	10.39	11.32	10.74	**11.66**
City in state	22.25	22.66	30.44	**31.82**
Family	71.54	67.00	72.73	**77.67**
Adjective to adverb	21.37	19.66	**23.79**	22.68
Opposite	14.41	13.42	**17.98**	17.00
Comparative	52.85	50.98	**67.04**	63.06
Superlative	31.02	28.16	**41.00**	39.22
Present participle	37.31	35.42	**38.35**	36.17
Nationality adjective	79.11	79.42	77.92	**79.61**
Past-tense	46.15	45.38	**48.27**	47.76
Plural	52.25	52.03	54.50	**55.56**
Plural-verbs	**42.76**	42.41	40.57	41.15

Table 3 Word semantic and syntactic accuracies. (%)

Percentage of paragraph (%)	CBOW		CBOW+coref	
	Semantic	Syntactic	Semantic	Syntactic
10	49.00	45.93	**52.52**	45.58
25	50.44	47.83	**52.70**	47.00
50	52.20	48.61	**54.03**	48.07
75	51.96	48.85	**54.20**	48.43
90	51.98	48.67	**53.84**	47.73
100	51.75	48.88	**53.65**	48.16

4.3 Word similarity

In this section, we will review the task of word similarity, then present our results.

4.3.1 Word similarity and evaluation method

This task is given two pairs of words and the experiment is to determine if they are similar, based on their corresponding vector representation. There are several human annotated benchmarks that are widely used to measure word similarity. These benchmarks are either human-rated or crowdsourcing-related. Similarity between two words can be calculated in many ways, but usually by the cosine distance of two corresponding vectors. Then, the embedding is scored by Spearman's rank correlation coefficient between the embedding ranking and the human rankings. The Spearman's coefficient is a number from -1 to $+1$, where a Spearman correlation of 0 indicates that the embedding ranking is different from the human ranking, while the closer it is to 1 means the closer it matches the human ranking. We use the tool provided by Faruqui and Dyer[11] to evaluate the word similarity.

4.3.2 Results for word similarity

Four word embeddings are evaluated together with their corresponding co-reference context versions. Table 4 shows the results with word similarity ranking. We can see that the overall quality of the dependency-based context embedding has increased after applying the co-reference resolved context. Overall, the task correlation using MEN, MTurk-287, MTurk-771, and Verb-143 has seen an increase in the correlation rankings. We find

Table 4 Results for word similarity scoring on 12 different data sets.

Task name	skip	skip+coref	CBOW	CBOW+coref	GloVe	GloVe+coref	dep	dep+coref
WS-353	0.6803	0.6772	0.6281	0.6283	0.5477	0.5338	0.5314	0.5264
WS-353-SIM	0.7477	0.7422	0.7212	0.7084	0.6516	0.6395	0.7139	0.7164
WS-353-REL	0.6147	0.6169	0.5378	0.5340	0.4927	0.4783	0.3594	0.3462
MC-30	0.7564	0.7422	0.7429	0.7112	0.6260	0.6267	0.7157	0.7108
RG–65	0.7309	0.7124	0.7263	0.7230	0.6618	0.6720	0.6412	0.6258
Rare-Word	0.4222	0.4166	0.4258	0.4177	0.3319	0.3253	0.3200	0.3220
MEN	0.7028	0.7036	0.6747	0.6763	0.6536	0.6545	0.5583	0.5659
MTurk-287	0.6717	0.6691	0.6662	0.6699	0.6191	0.6229	0.5623	0.5664
MTurk-771	0.6051	0.6026	0.5820	0.5881	0.5897	0.5884	0.5050	0.5194
YP-130	0.4150	0.4251	0.2729	0.2651	0.4307	0.3613	0.2022	0.2320
SimLex-999	0.2869	0.2855	0.2963	0.2912	0.2710	0.2671	0.3514	0.3569
Verb-143	0.3663	0.3827	0.3528	0.3446	0.3404	0.3474	0.4317	0.4350

that using the co-reference context has a larger positive effect on the dependency embedding compared to the other groups. But in general, the difference is minor. Another point is that using the co-reference context can help us find more pairs of words. Hence, in this case, this is an advantage for using co-reference embedding.

We aim to find how pronouns affect word embedding and which one is affected more using the co-reference context. In our experimental results, the performance improvements using the co-reference context are significant for word semantic analysis, but are not prominent for the syntactic analysis. These results are confirmed in Table 3, and also supported by Tables 2 and 4. Multiple diverse types of datasets are included in Tables 2 and 4. Some datasets belong to the semantic relation type, while others belong to the syntactic relation type. For example, the capitals of common countries are of the semantic type, where the adjective to adverb is of the syntactic type. Similarly, the improvement on the semantic datasets is more obvious than that on the syntactic data sets.

5 Related Work

The task of representation learning is to generalize where probability mass concentrates and to leverage the defect brought by the curse of dimensionality. Two main directions for addressing this problem are feature engineering and feature learning. Hand crafting features is considered labor-intensive while representation learning[14] learns dense vector with low space. This paper focuses on auto-generated feature extraction models, especially on the skip-gram neural embedding model[3] and other models derived from this.

The skip-gram model maximizes the probability that a word and its context belong to a document, while negative sampling adds a negative sample and maximizes the probability that a word and its context belong to the document, and at the same time the probability that negative samples do not belong to the document. Following this work, a dependency-based word embedding was introduced[8] that replaced the original linear bag-of-words context with the dependency-based context, making the context syntactically closer to the word. As these models can learn a very compact representation of words, the word co-reference resolution is not considered. In our example from the introduction, we demonstrated that these models do not handle these problems. In the work of Huang et al.upcite15, the quality of word representation was improved using the global context and ambiguous words were resolved with multiple word prototypes. The GloVe model is another embedding model that performs aggregated global word-word co-occurrence statistics from a corpus for embedding training. Some other embedding models include: Neural Network Language Model (NNLM)[16], Log-Bilinear Language (LBL) Model[17], C&W model[18], and the Character-level Neural Language Models[19–21].

But these models look at the corpus in general and do not consider how small things can make changes. They do not consider the fact that names are very important in some NLP tasks and focus on having a generally improved word embedding. The closest work that we can find is that by Adel and Schutze[22]. But their work uses the co-reference resolution to extract discontinuous linguistic units whereas our work uses the co-reference resolution to increase the quality of embedding. Our work generalizes how the co-reference resolved context

can be included into the context for embedding training.

6 Conclusion

Pronouns play a key role in written articles and may appear in large numbers. This paper focuses on investigating the effect these pronouns have on word embedding. We present a general method to include the co-reference resolved context in the context extracting phrase. We perform the experiments on four word embeddings and compare them to the embeddings with the co-reference resolved context. We experiment with two sets of embedding on the analogy task and all sets of embedding on word similarity. Experimental results show that the co-reference resolved context embedding outperforms the original context embedding by an average of 1.88% in semantic accuracy in the word analogy. This work has presented an interesting future of using both resolved and unresolved context simultaneously to train better embedding.

Acknowledgment

This work was supported by the National High-Tech Research and Development (863) Program (No. 2015AA015401), the National Natural Science Foundation of China (Nos. 61533018 and 61402220), the State Scholarship Fund of CSC (No. 201608430240), the Philosophy and Social Science Foundation of Hunan Province (No. 16YBA323), and the Scientific Research Fund of Hunan Provincial Education Department (Nos. 16C1378 and 14B153).

References

[1] Z. S. Harris, Distributional structure, *Word*, vol. 10, nos. 2&3, pp. 146–162, 1954.

[2] H. Rubenstein and J. B. Goodenough, Contextual correlates of synonymy, *Communications of the ACM*, vol. 8, no. 10, pp. 627–633, 1965.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

[4] Q. V. Le and T. Mikolov, Distributed representations of sentences and documents, arXiv preprint arXiv:1405.4053.

[5] T. Mikolov, S. Kombrink, L. Burget, and J. H. Cernocky, Extensions of recurrent neural network language model, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 5528–5531.

[6] T. Mikolov and G. Zweig, Context dependent recurrent neural network language model, in *IEEE Workshop on Spoken Language Technology (SLT)*, Miami, FL, USA, 2012, pp. 234–239.

[7] T. Mikolov, A. Deoras, D. Povey, and L. Burget, Strategies for training large scale neural network language models, in *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, HI, USA, 2011, pp. 196–201.

[8] O. Levy and Y. Goldberg, Dependency-based word embeddings, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Baltimore, MD, USA, 2014, pp. 302–308.

[9] J. Pennington, R. Socher, and C. Manning, GloVe: Global vectors for word representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543.

[10] M. Baroni, G. Dinu, and G. Kruszewski, Don't count, predict! A systematic comparison of context-counting vs. contextpredicting semantic vectors, in *Proceedings of the 52^{nd} Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Baltimore, MD, USA, 2014, pp. 238–247.

[11] M. Faruqui and C. Dyer, Community evaluation and exchange of word vectors at wordvectors.org, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics*, Baltimore, MD, USA, 2014, pp. 19–24.

[12] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky, Stanfords multi-pass sieve coreference resolution system at the conll-2011 shared task, in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, CONLL Shared Task 11*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 28–34.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, Distributed representations of words and phrases and their compositionality, in *Advances in Neural Information Processing Systems*, Lake Tahoe, CA, USA, 2013, pp. 3111–3119.

[14] Y. Bengio, A. C. Courville, and P. Vincent, Unsupervised feature learning and deep learning: A review and new perspectives, arXiv preprint arXiv:1206.5538v1.

[15] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, Improving word representations via global context and multiple word prototypes, in *Proceedings of the 50^{th} Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 873–882.

[16] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, A neural probabilistic language model, *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[17] A. Mnih and G. Hinton, Three new graphical models for statistical language modelling, in *Proceedings of the 24^{th} International Conference on Machine Learning, ICML 07*, ACM, New York, NY, USA, 2007, pp. 641–648.

[18] R. Collobert and J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in *Proceedings of the 25th International*

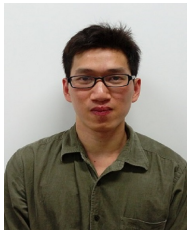
Conference on Machine Learning, ICML 08, ACM, New York, NY, USA, 2008, pp. 160–167.

[19] I. Sutskever, J. Martens, and G. E. Hinton, Generating text with recurrent neural networks, in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, Bellevue, WA, USA, 2011, pp. 1017–1024.

[20] A. Graves, Generating sequences with recurrent neural networks, arXiv preprint arXiv:1308.0850.

[21] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, Character aware neural language models, in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI Press, Phoenix, AZ, USA, 2016, pp. 2741–2749.

[22] H. Adel and H. Schtze, Using mined coreference chains as a resource for a semantic task, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, Doha, Qatar, 2014, pp. 1447–1452.



Tonglee Chung received the BS degree in computer science and technology from Jinan University, China, in 2012. He is currently a PhD candidate in the Department of Computer Science and Technology, Tsinghua University. His research interests include knowledge base construction and machine learning.

Bin Xu is an associate professor in Department of Computer Science and Technology of Tsinghua University. He obtained the PhD, master, and bachelor degrees from Tsinghua University in 2006, 1998, and 1996, respectively. He became an ACM professional member in 2009 and on IEEE member in 2007. His research interests include knowledge graph, semantic web, and service computing.

Yongbin Liu received the PhD degree from University of Science & Technology Beijing, China, in 2013. From 2013 to 2015, he was a post-doc research fellow in Tsinghua University. He is an associate professor in University of South China. His research interests include natural language processing and knowledge engineering.

Juanzi Li received the PhD degree from Tsinghua University, China in 2000. She is currently a professor at the Department of Computer Science and Technology in Tsinghua University. Her research interests are semantic web and semantic Web services, text mining, and knowledge discovery. She is a member of China Computer Federation (CCF) and Association for Computing Machinery (ACM).

Chunping Ouyang received the PhD degree from University of Science & Technology Beijing, China, in 2011. From 2014 to 2015, she was a visiting scholar at Tsinghua University. She is an associate professor of computer science at University of South China and supervisor of postgraduate. She has served as the program committee member of various international conferences and reviewer for various international journals. Her research interests include natural language processing and information retrieval.