Distant Supervision for Relation Extraction with Multi-sense Word Embedding

Sangha Nam, Kijong Han, Eun-kyung Kim and Key-Sun Choi
School of Computing, KAIST
Daejeon, Republic of Korea.
{nam.sangha, han0ah, kekeeo, kschoi}@kaist.ac.kr

Abstract

Distant supervision can automatically generate labeled data between a large-scale corpus and a knowledge base without utilizing human efforts. Therefore, many studies have used the distant supervision approach in relation extraction tasks. However, existing studies have a disadvantage in that they do not reflect the homograph in the word embedding used as an input of the relation extraction model. Thus, it can be seen that the relation extraction model learns without grasping the meaning of the word accurately. In this paper, we propose a relation extraction model with multi-sense word embedding. We learn multi-sense word embedding using a word sense disambiguation module. In addition, we use convolutional neural network and piecewise max pooling convolutional neural network relation extraction models that efficiently grasp key features in sentences. To evaluate the performance of the proposed model, two additional methods of word embedding were learned and compared. Accordingly, our method showed the highest performance among them.

1 Introduction

Relation extraction refers to the task of extracting the relation between two entities in a sentence. For example, a relation extraction system extracts '*Founder(Facebook, Mark Zuckerberg)*' from the sentence "*Mark Zuckerberg is the founder of Facebook*". In recent years, the importance of knowledge bases has emerged, and studies for constructing large-scale knowledge bases such as DBpedia, YAGO, and Wikidata are actively underway. Furthermore, the research on extracting knowledge from web-scale corpus is also underway. However, since many studies use machine learning to design a relation extraction system, there is a high-cost problem in generating a large amount of supervised training data. To solve this problem, the distant supervision assumption is introduced in this paper (Mintz *et al.*, 2009). The distant supervision assumption means, "*If two entities are linked with a certain relation in the knowledge base and there is a collected sentence that contains both entities from the corpus, then the collected sentences may describe the certain relation between the two entities.*" Figure 1 is an example of automatically collected labeled data using the distant supervision assumption.

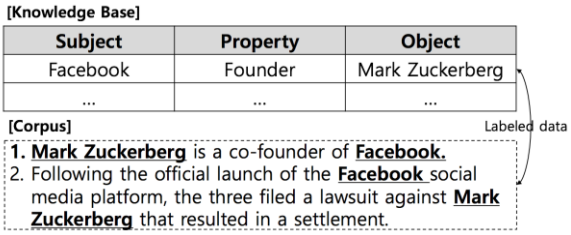

Figure 1. Example of labeled data collection based on distant supervision

The distant supervision method is relatively efficient in that it automatically generates training/labeled data between a large corpus and a large knowledge base, but the veracity of the labeled data is sometimes ambiguous. As shown in Figure 1, among the collected sentences that contain both '*Facebook*' and '*Mark Zuckerberg*', the first sentence means that Mark Zuckerberg is a founder of Facebook, but the second sentence does not. Various studies (Riedel *et al.*, 2010; Hoffmann *et al.*, 2011; Surdeanu *et al.*, 2012) have been introduced to solve this problem. However, they use traditional natural language processing (NLP) features such as part of speech (POS) tagging and dependency tree, so the errors occurring in NLP tools propagate to the relation extraction system. Therefore, these papers (Kim, 2014; Zeng *et al.*, 2014) proposed a relation extraction system that used word embedding and deep neural network (DNN) approaches without the above NLP features, and showed improved performance than previous studies. Especially, the piecewise max pooling convolution neural network (PCNN) model introduced in (Zeng *et al.*, 2015) transforms the convolution neural network (CNN) model into a form more suitable for relation extraction task.

However, these studies have a disadvantage in not reflecting the sense of words in word embedding. For example, the word '*bow*' could be divided into various meanings such as '*baU – greeting*' and '*boU – archer's weapon*'. Therefore, if a relation extraction model is learned with lexical ambiguity, it may result in not properly reflecting the characteristics of the homograph. Thus, it is necessary to apply multi-sense word embedding to the relation extraction model. However, to the best of our knowledge, there are no studies applying multi-sense word embedding to relation extraction models.

In this paper, we introduce a distant supervision relation extraction model with multi-sense word embedding. We use two relation extraction models, CNN proposed in (Kim, 2014) and PCNN proposed in (Zeng *et al.*, 2015). To learn the multi-sense word embedding, we use the results of the word sense disambiguation (WSD) module and Skip-gram algorithm. To demonstrate the superiority of our method, we compared the relation extraction performances of two other word embedding models. The first is the most common word-token-based word embedding, and the second is the morpheme-based word embedding. In chapter 4, we present the experimental results of learning and evaluation of these models based on Korean Wikipedia and K-Box, which extended knowledge base on Korean DBpedia.

2 Related Work

2.1 Skip-gram Model

Word embedding is a way of expressing words in real-valued vectors, and expresses the meaning of a word on the vector space. Thus, it is easy to grasp the semantic similarity between words by a simple vector operation, and therefore, it is widely used in various NLP fields. The skip-gram model, which is type of word embedding learning method, learns by predicting words that appear around the target word. The skip-gram model proceeds to maximize the following objective function.

$$J(\theta) = \sum_{(w_t,c_t)\in D+} \sum_{c\in c_t} logP(D=1\,|v(w_t),v(c)) + \sum_{(w_t,c_t')\in D-} \sum_{c'\in c_t'} logP(D=0\,|v(w_t),v(c'))$$

w_t is a target word and c_t stands for the word actually appearing around w_t in the corpus, and c_t' are randomly selected words that do not appear around w_t. That is, the learning is performed in such a manner as to maximize the probability of predicting words actually appearing around a target word and the probability of not predicting words that did not actually appear.

2.2 PCNN Relation Extraction Model

CNN is a deep neural network that shows excellent performance in image classification and sentiment classification. One of the features and advantages of CNN is that it efficiently finds key features in input data. Accordingly, the authors in (Kim, 2014; Zeng *et al.*, 2014) proposed a relation extraction model using CNN. In (Zeng *et al.*, 2014), the authors suggest the position embedding concept and adding it to the input vector of their CNN relation extraction model, and then the performance is improved. Position embedding is the embedding of the relative distance between two entity and non-entity words in a sentence as an n-dimensional vector. For example, as shown in Figure 3, the word '*co-founder*' is three words away from the '*Mark Zuckerberg*' entity and two words away from the '*Facebook*' entity. This relative distance is embedded into an n-dimensional vector to create the position embedding, and the value is used as part of the input vector of model learning.

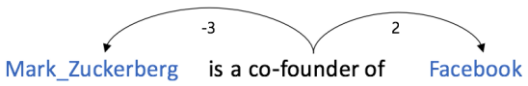

Figure 3. Example of the relative distance of position embedding

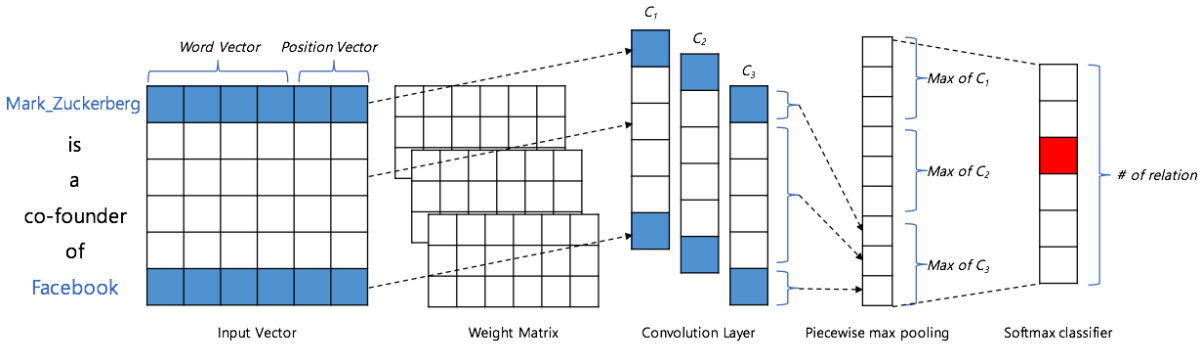

Figure 2. Architecture of PCNN

PCNN is an extended CNN model proposed in (Zeng *et al.*, 2015). The structure of PCNN is shown in Figure 2. The entire structure is made up of input vectors, three convolution layers, piecewise max-pooling layer, and softmax output layer. The input vector consists of a word vector and a position vector. The major difference is that extends the single max-pooling layer to the piecewise max-pooling layer. In CNN, max pooling is the method of extracting the largest value, i.e., the most important feature, in the output matrix of the convolution layer. However, it is difficult to grasp the key features required for relation extraction by selecting only one maximum value among the convolution layer result values in the single max-pooling layer. To solve these weaknesses, PCNN proposed a piecewise max-pooling layer by dividing the single max-pooling layer into three. Since the sentence used in relation extraction always contains two entities, it is possible to divide the sentence into three subunits based on two entities, and then the maximum value is extracted for each subunit in the piecewise max-pooling layer.

3 Methodology

In this paper, we propose a relation extraction model using multi-sense word embedding. We use CNN and PCNN for the relation extraction model, and generate multi-sense word embedding using the WSD module.

The structure of our relation extraction system is as shown in Figure 4, and it consists largely of the word embedding and distant supervision relation extraction model. First, we take the corpus as input and perform WSD module. Next, entity-padding tokenization is performed as described in Section 3.1. Next, the multi-sense word embedding is learned by the skip-gram algorithm, so that the tokens with the sense number have their own embedding vectors. In this way, the same form of lexical token has different embedding vectors based on the sense number.

Distant supervision is performed between Knowledge Base and Corpus, and the collected labeled data are word sense disambiguated and tokenized in the same manner. Then this data is divided into two groups—one for learning and the other for evaluation.

3.1 Multi-sense Word Embedding

In general, word embedding is a method of dividing the input corpus into word tokens and then mapping tokens with similar meaning onto similar vector spaces. In English, a token is usually generated in word units. However, Korean language is not as good as English when the word embedding is generated on a word token due to the plurality of elements constituting a word such as postposition, ending, and suffix. Therefore, when learning word embedding in Korean, a token is formed by a stem unit, and a POS tag is sometimes used as a constituent element of the token. The advantage of using a POS tag in learning of word embedding is that it can be divided into whether the same lexical word is used as a verb or as a noun. For example, in Korean, the word '*Ga-Ji*' can be used as a noun to mean '*branch*' or as a verb to mean '*get*'. Moreover, as an example in English, the word '*wind*' can be used as '*movement of air*' for nouns and '*twist*' for verbs. Therefore, when learning word embedding, it is effective to use POS tags together to construct a token because some ambiguity could be resolved.

However, there is a problem in that common word embedding does not reflect the actual meaning of words, which is the same in Korean as well as English. For example, the word '*apple*' is used both as a fruit and as a company. As mentioned earlier in Chapter 2, word embedding is based on what the surrounding words appear to be. The words around '*apple-fruit*' and the words around '*apple-company*' are definitely different, but all these words appear around the word '*apple*', so the word '*apple*' has only one n-dimensional real-valued vector that cannot distinguish between

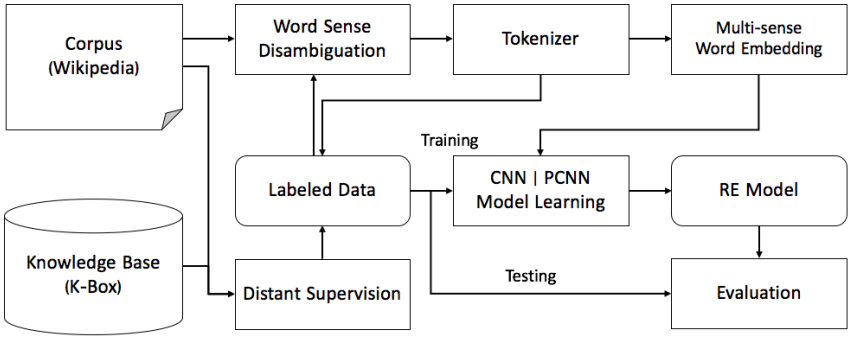

Figure 4. Architecture of relation extraction system with multi-sense word embedding

'*apple-fruit*' and '*apple-company*'. Thus, the triangle inequality problem (Neelakantan *et al.*, 2015) can occur.

$$distance(a,c) \leq distance(a,b) + distance(b,c)$$

For example, there is a problem that the distance between '(a) *pollen* – (c) *refinery*' is smaller than the sum of the distances between '(a) *pollen* – (b) *plant*' and '(c) *refinery* – (b) *plant*'. In other words, the similarity between the two words '*pollen*' and '*refinery*' is closer to the actual semantic distance centered on the homonym of '*plant*'. To solve this problem, several papers (Neelakantan *et al.*, 2015; Rothe and Schütze, 2015) have been published that learn word embedding by the actual meaning of words using a method is called multi-sense word embedding.

We learn multi-sense word embedding using a WSD module to distinguish the meaning of words in advance. Our WSD module is based on the unsupervised learning approach and uses the Markov Random Field (MRF) algorithm which resolves the ambiguity based on the semantic category of CoreNet (Choi *et al.*, 2004). In MRF, the node is composed of common noun, verb, and adjective, and the edge between the nodes is set as long as the distance is only one on the dependency path, in a similar way to this paper (Chaplot *et al.*, 2015).

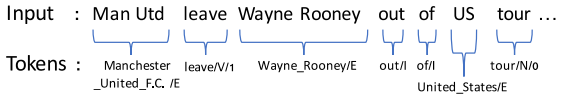

Figure 5. Example of Tokenization

The tokenization example for the input sentence is shown in Figure 5. The second word '*leave*' is tokenized with a POS tag and a sense number. In addition, to make a word embedding suitable for relation extraction, the multiword entity was grouped into one token. As shown in Figure 5, '*Man Utd*' and '*Wayne Rooney*', a multiword entity, was bundled into a single token, and solved the entity disambiguation problem. Even if an entity consists of several words, learning to have a single word embedding value is proper for designing a word embedding and relation extraction model. We use personal entity tags in Wikipedia for entity linking as shown in Figure 6.

Facebook is an American for-profit corporation and an online social media and social networking service based in Menlo Park, California. The Facebook website was launched on February 4, 2004, by Mark Zuckerberg,

Figure 6. Example of Multiword Entity in Wikipedia

These blue entities, such as '*for-profit corporation*', '*social media*', and '*social networking service*', are hand-tagged by Wikipedia content writers, so they are very accurate.

3.2 Relation Extraction Model

We use CNN and PCNN relation extraction models. The input representation consists of a 100-dimensional word vector and a 10-dimensional position vector. Three convolution layers were constructed and the weight matrix size was 3 × 110, and the stride is one. CNN model is implemented as a single max-pooling layer and PCNN model is implemented as a piecewise max-pooling layer. The softmax layer is sized according to the relation number of the classification.

4 Evaluation

4.1 Data

For the experiment, we used 6,941,760 sentences of Korean Wikipedia (2017. 07. 01) and K-box. K-Box is a knowledge base that extends triple to Korean DBpedia, and the added triple is a conversion of Korean local property into ontological property. For example, the conversion of a Korean property such as '*prop-ko:chul-saeng-ji*' into a '*dbo:birthPlace*'. The mapping table is created manually by three human experts. Through distant supervision, 358,464 labeled data were collected on 451 properties in all, but many properties were long tail problems with a small amount of collected data. In the multi-class classifier model, since learning does not proceed properly if there are few data per class, we used total 200,323 labeled data of 68 properties based on the number of collected data, which is 1000 or more per class.

4.2 Evaluation Results

To demonstrate the excellence of our proposed method, three types of word embedding have been learned. The first is learning by tokenization in word unit (Word), the second is tokenization by morpheme unit and POS tag (+POS), and the third is tokenization by morpheme unit, POS tag, and word sense (++WSD). All of these types of learning proceeded with the same parameters; 100-dimension, 5 window sizes, 1 minimum word count.

As given in Table 1, the result of multi-sense word embedding clusters the words in a sense-specific manner. In addition, since we apply multiword entity embedding, we can see that the multiword entity is learned by one embedding vector, and the similar words are also meaningful.

Token	Word	Similar Words
+POS	Si-Jang	invest, distribution, profit, export, assets, conglomerate, sales, import, industry, price
	Sa-Gwa	ask, apology, sorry, condolences, pass, envelope, report, complain, explanation, comment
++WSD	Si-Jang - Market	industry, business, competitiveness, small businesses, enterprise, investment, antioxidant, finance
	Si-Jang - Mayor	superintendent of education, self-government director, Park Soonja, The 5th Local Elections in Korea
	Sa-Gwa - Apology	apology, pass, accusation, sorry, morning star, :'(
	Sa-Gwa - Apple	fruit, pea, chestnut, apricot, walnut, grape, nut products, poison ivy
Entity	UN	United Nations, European Community, North Atlantic Treaty Organization, League of Nations, Security Assurance

Table 1: Similar words of 'Si-Jang' and 'Sa-Gwa' by word embedding. 'Si-Jang' is a Korean word, and it is mainly used for market or mayor. 'Sa-Gwa' is also a Korean word, and it is mainly used for apology or apple. All of the similar words are written by translating Korean words into English.

We perform the held-out evaluation of the relation extraction model using the multi-sense word embedding. Held-out evaluation is a method for measuring precision, recall, F1-score by dividing the collected data in half, and one is used for learning and the other for evaluation. The evaluation results are shown in Table 2. To verify the effectiveness of our method, we used three different embedding models, as mentioned above, as inputs of the CNN/PCNN models, and measured the performance. The hyper-parameters settings for two models are as follows; both models were set to ReLU activation and 1 drop-out, but CNN use Adadelta optimizer and PCNN use Adam optimizer.

Owing to the evaluation, both models showed better performance of using morpheme embedding (+POS) than word embedding (Word), and the performance of sense embedding (++WSD) is also improved than morpheme embedding (+POS). Additionally, the performance of the CNN model was higher than that of the PCNN model because, in Korean, the position of two entities in the sentence often appears at the top of the sentence and the two entities are often placed consecutively.

Model	Embedding	Precision	Recall	F1-score
CNN	Word	0.5537	0.3506	0.4275
	+POS	0.5315	0.4279	0.4739
	++WSD	**0.5921**	**0.5039**	**0.5443**
PCNN	WSD	0.457	0.3251	0.3799
	+POS	0.4555	0.3472	0.394
	++WSD	0.4529	0.3713	0.4081

Table 2: Performance of relation extraction model by word embedding

5 Conclusion

In this paper, we propose a method for improving the performance of a distant supervision relation extraction model using multi-sense word embedding, and experimentally evaluated two relation extraction models based on CNN and PCNN. In addition, we used entity-padding word embedding, which bundles multi-word entity into a single token, when generating word embedding. Accordingly, it was confirmed that the multi-sense word embedding improves the performance of the relation extraction model.

In the future, we plan to apply the convolutional RNN model, which is a combined model of CNN and recurrent neural network (RNN), to the relation extraction task. We will also study the method of removal of error data, which is one of the problems when collecting distant supervised training data.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (2013-0-00109, WiseKB: Big data based self-evolving knowledge base and reasoning platform)

References

Chaplot, D. S., Bhattacharyya, P., & Paranjape, A. 2015. *Unsupervised Word Sense Disambiguation Using Markov Random Field and Dependency Parser*. In Proceedings of AAAI, pages 2217-2223.

Choi, K. S., Bae, H. S., Kang, W., Lee, J., Kim, E., Kim, H., ... & Shin, H. 2004. *Korean-Chinese-Japanese*

Multilingual Wordnet with Shared Semantic Hierarchy. In Proceedings of LREC, pages 1131-1134.

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., & Weld, D. S. 2011. *Knowledge-based weak supervision for information extraction of overlapping relations.* In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, pages 541-550.

Kim, Y. 2014. *Convolutional neural networks for sentence classification.* In Proceedings of EMNLP.

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

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. 2009. *Distant supervision for relation extraction without labeled data*. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, pages 1003-1011.

Neelakantan, A., Shankar, J., Passos, A., & McCallum, A. 2015. *Efficient non-parametric estimation of multiple embeddings per word in vector space*. arXiv preprint arXiv:1504.06654.

Riedel, S., Yao, L., & McCallum, A. 2010. *Modeling relations and their mentions without labeled text.* Machine learning and knowledge discovery in databases, pages 148-163.

Rothe, S., & Schütze, H. 2015. *Autoextend: Extending word embeddings to embeddings for synsets and lexemes.* arXiv preprint arXiv:1507.01127.

Surdeanu, M., Tibshirani, J., Nallapati, R., & Manning, C. D. 2012. *Multi-instance multi-label learning for relation extraction.* In Proceedings of EMNLP, Association for Computational Linguistics, pages 455-465.

Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J. 2014. *Relation Classification via Convolutional Deep Neural Network.* In Proceedings of COLING, pages 2335-2344.

Zeng, D., Liu, K., Chen, Y., & Zhao, J. 2015. *Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks.* In Proceedings of EMNLP, pages 1753-1762.