Incorporating Global Contexts into Sentence Embedding for Relational Extraction at the Paragraph Level with Distant Supervision

Eun-kyung Kim, Key-Sun Choi
Semantic Web Research Center, School of Computing, KAIST
291 Daehak-ro, Yuseong-gu, Daejeon, South Korea
{kekeeo, kschoi}@kaist.ac.kr

Abstract

The increased demand for structured knowledge has created considerable interest in relation extraction (RE) from large collections of documents. In particular, distant supervision can be used for RE without manual annotation costs. Nevertheless, this paradigm only extracts relations from individual sentences that contain two target entities. This paper explores the incorporation of global contexts derived from paragraph-into-sentence embedding as a means of compensating for the shortage of training data in distantly supervised RE. Experiments on RE from Korean Wikipedia show that the presented approach can learn an exact RE from sentences (including grammatically incoherent sentences) without syntactic parsing.

Keywords: Relation Extraction, Sentence Embedding, Pro-drop Languages

1. Introduction

As the demand for structured knowledge has increased, considerable interest has emerged in relation extraction (RE) from large collections of documents written in natural language. In particular, with "distant supervision" (DS) (Mintz et al., 2009; Hoffmann et al., 2011; Riedel et al., 2013), it is possible to extract the relationships between pairs of entities without human manual annotation using a knowledge base (KB); this heuristically aligns entities in texts to a given KB and then uses this alignment to train an RE system.

Although the DS strategy is a more effective method of automatically labeling training data than directly supervised labeling, DS-based approaches can extract only relations that are limited to a "single complete sentence" that contains two target entities. This makes it difficult to obtain both the subjects and object entities that participate in the KB in a single sentence, particularly in *null subject (or object) languages* such as Korean, Japanese, Arabic, and Swedish, that can leave the subject of a sentence unexpressed, unlike English which allows neither. It is also difficult to utilize DS-based approaches for English data when sentences have an informal, grammatically incoherent style, such as the style popularly used on Twitter, in discharge summaries of clinical texts (Marsh and Sager, 1982), or in a text shortened to bulleted lists in a Wikipedia article.

This point can be illustrated by considering the examples in Figure 1. S_1 contains a subject, object, and predicate, whereas the subject is omitted in S_2 because it is obvious in adjacent sentences in Korean, resulting in differences between the same sentence written in Korean and in English. Therein, we know S_2 is obviously a positive example for tuple `founderOf`(*Steve Jobs*, *Apple Inc.*), but we cannot label the training instance S_2 according to the traditional paradigm of an existing DS-based approach.

We propose a novel approach that performs RE across sentences, at the paragraph-level, and does not require labeled data. The proposed method builds upon sentence embedding with global context constraints by spanning multiple

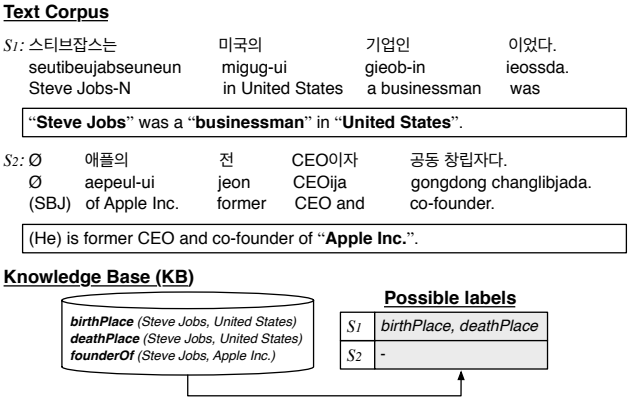

Figure 1: The English sentences are both correct translations of each Korean sentence. The entities in the sentence are marked in boldface with parenthesized boundaries.

sentences, which is useful for estimating omitted subjects and predicting relations. First, we specifically perform novel zero subject resolution with the entity-relation-based graph analysis by applying the centrality measure. This allows us to learn RE models for informal sentences and has the advantage of compensating for a shortage of training data in the DS-based approach in a DS-based approach to null subject languages. Then, we try to capture the discriminative context features of each document type, such as the specific logical pattern to the relational flow of text within a paragraph, to support sentence embedding.

Our work differs from previous related works in two ways: (1) we propose a method of RE at the paragraph-level—i.e., from a collection of multiple sentences— rather than extracting information from an independent single sentence; (2) our approach, which builds upon the sentence embedding, is more effective for language-independent extraction because it avoids high-level natural language processing (NLP) tools. Therefore, the present approach can be generally used for RE, even in languages for which NLP tools are lacking.

2. Related Work

We often encounter a lack of explicitly annotated text in RE, instead finding richly structured KBs such as DBpedia (Bizer et al., 2009) or Freebase (Bollacker et al., 2008), which has raised significant interest in learning RE using DS. Many DS-based approaches (Hoffmann et al., 2011; Roller and Stevenson, 2014; Tsai and Roth, 2016; Craven and Kumlien, 1999; Mintz et al., 2009) use simple but effective heuristics to align existing facts with unlabeled text. This automatically generated labeled text can be used as training data for supervised learners. Our work was inspired by Mintz et al. (2009), who adopted the Freebase for the distant supervision of the Wikipedia corpus. Unlike existing methods, we performed RE across sentences at the paragraph-level by extending the possibility of labeling incomplete sentences that were unavailable in the traditional DS-based approach. To the best of our knowledge, ours is the first DS-based approach to solve the problem of data sparseness by applying DS to the RE of informal sentences and alleviating DS assumptions.

Quirk and Poon (2016) introduced the RE method in two adjacent sentences using the DS approach. Peng et al. (2017) explored a framework for cross-sentence n-ary RE based on graph long short-term memory networks; they used a graph formula to provide a unified method of integrating various intra- and inter-sentential dependencies such as sequential, syntactic, and discourse relationships. According to their experiments on biomedical domains, use of RE beyond sentence boundaries can yield much more knowledge. In this context, we intend to find more information by spanning multiple sentences. While they are based on the various linguistic analyses, our proposed method can be differentiated by using contexts without syntactic information.

3. Relation Extraction at the Paragraph-level

We define our task as follows: Given a sentence s' that is a complemented form of an informal (e.g. subject-less) sentence s with marked entities e_1 and e_2 and a set of relations $R = \{r_1, ..., r_n\}$, we formulate the task of identifying the semantic relation as a standard classification problem as follows:

$$f : (P, E, L) \rightarrow R, \quad (1)$$

where P is the set of all paragraphs, a paragraph $p \in P$ is the set of contiguous sentences $\{s'_1, s'_2, \ldots, s'_m\}$, E is the set of entity pairs, and L denotes the set of relation flows. A relation flow $l \in l$ is a tuple $(\overleftarrow{s'}, \overrightarrow{s'})$ in which $\overleftarrow{s'} = \{r_1, r_2, ...\}$ is the set of labeled relation mentions in which the preceding sentences are $(s'_1, ..., s'_{i-1})$ and $\overrightarrow{s'}$ is the set of labeled relation mentions in which the succeeding sentences are $(s'_{i+1}, ..., s'_m)$ with a given target sentence s'_i. Our training objective is to learn a joint representation of the sentences and the logical pattern of the relation flow of text within paragraphs such that a regression layer can predict the correct label. We propose an architecture that learns sentence embedding after compensating sentences with zero subject resolution.

S_1	[Steve Jobs]$_{e_1}$ was a [Businessman]$_{e_2}$ in [United States]$_{e_3}$.
S_2	ø$_1$ Former [CEO]$_{e_4}$ and co-founder of [Apple Inc.]$_{e_5}$.
S_3	On October 5, 2011, ø$_2$ died of [Pancreatic cancer]$_{e_6}$.

Table 1: Entity-tagged sentences taken from the first paragraph of the "Steve Jobs" article in the Korean edition of Wikipedia.

3.1. Zero Subject Resolution using Graph Analysis of a Paragraph

The basic idea of our zero subject (entity) prediction is to perform tasks by finding the central entity being described within a paragraph without parsing. This prediction task allows us to apply our method to many languages in which NLP tools are lacking. We hypothesize that the paragraph consists of contiguous sentences that describe the central entity. Given an unlabeled textual training corpus (Φ = Wikipedia) and the supervision KB (Ψ = DBpedia), we first identify all paragraphs in Φ and entities ($\in \Psi$) in the sentence. For example, for S_1, S_2, and S_3 in Table 1, we use WikiLinks[1] to identify six DBpedia entities in total.

When entities in every sentence of a given paragraph are identified, the entity graph $\mathbb{G}$ in the paragraph is constructed based on the relation tuple in Ψ between a pair of entities that appear in the paragraph. Then, the center node is selected in $\mathbb{G}$ based on the degree centrality (Wasserman and Faust, 1994) for assigning the latent subject entity beyond the sentence boundary. Centrality is important if the entity links to many other entities with one or multiple links to other entities in $\mathbb{G}$. The example entity graph G generated with e_1–e_6 is shown in Figure 2; (a) represents a tuple in the given Ψ between a pair of entities that appear together in the paragraph and (b) represents a digraph of the tuples shown in (a), where "Steve Jobs" is selected as the pivot by the out-degree centrality measure.

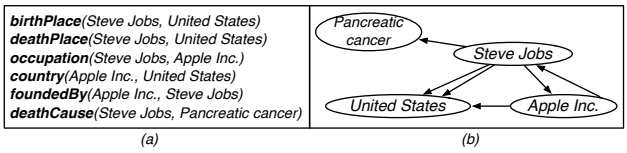

Figure 2: Example graph with given relation tuples between a pair of entities in the paragraph.

In this paper, the selected center entity is used to resolve zero subjects. Accordingly, a pair of entities that appear together in a single sentence, or head a pivoting entity and appear within a paragraph, is considered a potential relation instance. In the case of sentences S_2 and S_3 in Table 1, the concealed subjects, ø$_1$ and ø$_2$, both becomes "Steve Jobs" and provide an opportunity for acquiring possible labeled instances via heuristic alignments such as `founderOf`(*Steve Jobs*, *Apple Inc.*) and

[1] https://en.wikipedia.org/wiki/Help:Link

deathCause(*Steve Jobs*, *Pancreatic cancer*). Neither sentence explicitly states that "Steve Jobs" has such relationships, but have become useful for learning at training time by our extended model. For example, a sentence compensated by a pivot entity is syntactically incomplete, but we may derive a relatively large weight for the context feature associated with `founderOf` such as "former CEO and co-founder of."

At this stage, the only context features we use from s' are the words themselves. The vector representation of these words can be obtained using the Paragraph2Vec framework proposed by (Le and Mikolov, 2014), which maps each word to a vector and then uses a vector to represent all the words in the context window and thus predict the vector representation of the next word. The basic idea behind this method is to use an additional paragraph token (that maps to a vector space using a different matrix from that used to map the word) from the previous sentence in the document in the context window. Then, using the embedding matrix $\mathbf{E_{sen}} \in R^{D \times |V|}$ where D is the dimension of embedded words and $|V|$ is the dimension of the word vocabulary, we can obtain the embedding of the word. All words were randomly initialized and then updated during training.

3.2. Relational Flow Generation

Through the background of incorporating a global context into sentence embedding, the important intuition in our proposed model is understanding the whole paragraph as a single flow document. In this paper, we use the intuitive concept that if the semantic flow of a paragraph can be grasped, the relation type with which to classify the target sentence can be more clearly determined by the relation type with the preceding and succeeding sentences. For this, our auxiliary task is to determine the sequence of how preceding and succeeding sentences are classified into their respective relation types. Figure 3 shows an example of each paragraph that consists of contiguous sentences for two different types of entity. When there are two types of entities—in this case "baseball player" and "president"—it is possible to use a pattern in which there is no relation "party" (a type of relationship that points to a group of politically organized people) in the baseball player paragraph, and "team" and "position" relations are found that are very close to one another. According to this, for all three sentences (S_1–S_3 in Table 1), relationship flow of S_1 is ($\emptyset$, {*founderOf*, *deathCause*}), that of S_2 is ({*birthPlace*, *deathPlace*}, {*deathCause*}), and that of S_3 is ({*birthPlace*, *deathPlace*, *founderOf*}, $\emptyset$). We embed this relational flow, thereby aiming to learn continuous representations of it in vector space, similarly to embedding of words. Thus, we can also represent each element, i.e. the preceding and succeeding sequences of the relational flow, as two one-hot vectors of the K-dimension, where K is equal to the amount of relational flow. We then use the matrixes $\mathbf{E}^{\leftarrow}_{\mathbf{flow}} \in R^{D \times K}$ and $\mathbf{E}^{\rightarrow}_{\mathbf{flow}} \in R^{D \times K}$ to obtain its embedding.

In succession, we directly concatenate the sentence vector $\mathbf{E_{sen}}$ and the relational flow vectors $\mathbf{E}^{\leftarrow}_{\mathbf{flow}}$ and $\mathbf{E}^{\rightarrow}_{\mathbf{flow}}$ to form the final feature vector. This results in low-dimensional sentence embedding where semantically woven sentences and the relation flows of paragraphs reside in the same part of the space that presents the semantic relationship. We use this vector to train the machine learning algorithm and classify relationships.

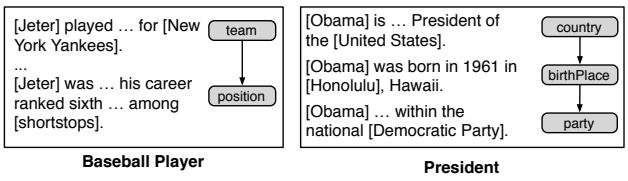

Figure 3: Relation sequence within a paragraph.

4. Experiments

We evaluated the performance of our proposed method by performing training and testing using the Korean version of Wikipedia as the textual corpus, specifically a snapshot from December 2016[2]. We used DBpedia to supervise background knowledge, which was a large KB of entities and relationships. As DBpedia provides tuple downloads in multiple languages[3], it was advantageous to build an efficient RE model for Koreans. KBs in non-Latin languages are relatively smaller than the English Freebase and DBpedia; our procedure used entities and tuples from DBpedia to provide relationship instances.

4.1. Implementation details

Distantly supervised RE can be viewed as a two-step process. This process (A) detects entities of interest and (B) determines the relationship between the possible set of entities. In this paper, we concentrate on the relationships between two entities, i.e., Step B. We processed the Wikipedia text using the following steps. (1) First, paragraphs are extracted from the article where a paragraph consists of two or more consecutive sentences that are separated by blank lines or different section names. (2) Second, the entities of sentences are identified using WikiLinks. In practice, an alternative entity recognition system may be required because the amount of text linked by WikiLinks is relatively small; however, that endeavor is beyond the scope of this study. (3) Third, central entities are selected from each paragraph by calculating the out-degree centrality based on the network model of the entity graph using the DBpedia tuple. (4) Fourth, sentences whose entity scope is recognized are tokenized. (5) Fifth, the pivot entity is employed to supplement the sentence and collect heuristically aligned data for the RE based on distant supervision. (6) Sixth, these labeled data are leveraged to construct sentence embedding, relation flow embedding, and finally to generate a single concentrated feature vector. (7) Finally, the RE model is trained with the feature vector to maximize the log probability of the correct relationship type. We converted each sentence into a word-level matrix in which each row was a sentence vector extracted from our model. Sentence vectors were learned from the Distributed Memory version of the Paragraph Vector (PV-DM) algorithm using training data

[2]https://dumps.wikimedia.org/kowiki/

[3]http://wiki.dbpedia.org/downloads-2016-04

to automatically learn and classify relationships into one of the 240 relation types in our evaluation dataset. PV-DM is an extension to Word2Vec (Mikolov et al., 2013) for learning document embeddings that was first applied to train using the entire corpus completely unsupervised.

We did not tune the initial learning rate (α) and minimum learning rate (α_{min}), and used the following values for all experiments: α = .025 and α_{min} = .002. The learning rate decreased linearly in each epoch from the initial rate to the minimum rate. We used the unchanged parameter min count (β) that represents the minimum frequency for times that a token must appear to be included in the Paragraph2Vec model's vocabulary. Our model set this as β = 1 to ensure that we treated all tokens in the context as meaningful and used them to train. We have optimized the embedding vector size (=400) and we used window sizes (=5) for the left and right fixed context windows. We ran an experiment with 10 epochs as the number of training iterations. All PV-DM training was carried out using the Gensim[4] library in Python. The next step was using a multiclass logistic regression classifier that was optimized using L-BFGS given the sentence embeddings inferred from the PV-DM model. Once the model had been trained, each sentence in the test dataset could be directly inferred.

4.2. Results Analysis of Extended Labeling

The original DS-based RE corresponds to a single sentence that contains two entities, but we extended this in this paper to tasks for two entities in a paragraph. We have made two extensions to the automated labeling schema, as described in Table 2. **Non-Extended** denotes the labeling results of two entities in a sentence according to the existing distant supervision paradigm. **Extended:Title** and **Extended:Pivot** are extensions of the label rather than Non-Extended. **Extended:Title** interprets the title of the Wikipedia document as the head entity because the title is the protagonist in the document, whereas **Extended:Pivot** represents the extension of the central pivoting entity in the paragraph as the subject entity, i.e. the proposed approach. Table 2 shows the proportion of judged documents for 50 sample documents and the precision—the proportion of relevant labeled sentences for RE—among that set.

It is clear from this table that the **Extended:Pivot** run achieved a higher or similar precision for the judged documents that it returned, but returned larger relevant labeled sentences (i.e. Positive Labels), and hence achieved a higher recall@R score, where R is the number of relevant documents in the collection. The **Extended:Title** method can also raise the precision and recall compared to the default DS paradigm in Wikipedia, but this is difficult to scale to a web-scale without a document title.

4.3. Held-out Evaluation for RE

We evaluated our RE model as a "held-out" evaluation. Such an evaluation is conducted automatically by withholding half of the DBpedia relationship knowledge during training and comparing the newly discovered relationship instances against the withheld data. The goal of automatic evaluation focuses on the accuracies of relation labels for each entity pair instead of the accuracies of the relation labels for each instance. We compared our model with the part-of-speech (POS) tag feature as a baseline that relies on the POS tag sequences of sentences for classification. Table 3 shows the results for the baseline for comparison with our algorithm. The best result was achieved using sentence embedding with relational flow, which led to an F1-measure of 59%. Although there is much room for improvement in precision and recall, our results indicated that it could be useful for extracting the relationship with small amounts of labeled data without advanced NLP tools such as a parser.

[4]https://radimrehurek.com/gensim/

	Total Labels	Sample Analysis (50 Doc.)	
		Positive Labels	Precision
Non-Extended	159,148	172	0.7257
Extended:Title	413,104	331	0.7405
Extended:Pivot	582,596	481	0.7527

Table 2: The corpus statistics before and after labeling extensions based on the distant supervision paradigm.

Features		Precision	Recall	F1-score
	Baseline	0.41	0.14	0.2087
$\mathcal{S}$	Non-Extended	0.45	0.39	0.4179
	Extended	0.53	0.50	0.5146
$\mathcal{S}+\mathcal{F}$	Non-Extended	0.59	0.54	0.5639
	Extended	0.60	0.59	0.5950

Table 3: Best F1-score measures with Precision and Recall for different feature sets, where $\mathcal{S}$ denotes the "sentence embedding" and $\mathcal{F}$ denotes the "relation flow embedding," by the incremental embedding of features compared with the POS-baseline.

5. Conclusion

In this paper, we focused on the distant supervision paradigm and proceeded to RE from passages that did not contain both of the entities that are expected to participate in a relation. We showed that it was possible to use a DS-based model that does not require labeling to represent the contexts of sentences and the surrounding relationship mentions to enable relation classification at the paragraph level. Experiments on Korean Wikipedia were conducted and showed the model's effectiveness in practical use. In future research, we intend to implement our technique on a much larger scale and with a more refined set of relation classifications. Alternatively, we may leverage cross-lingual joint techniques to transfer knowledge from other languages and to include joint learning with entity linking.

Acknowledgments

This work was supported by the Industrial Strategic technology development program (10072064, Development of Novel Artificial Intelligence Technologies To Assist Imaging Diagnosis of Pulmonary, Hepatic, and Cardiac Diseases and Their Integration into Commercial Clinical PACS Platforms) funded by the Ministry of Trade Industry and Energy (MI, Korea) and supported by the Bio & Medical

Technology Development Program of the NRF funded by the Korean government, Ministry of Science and ICT(NRF-2015M3A9A7029725)

6. Bibliographical References

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). Dbpedia – a crystallization point for the web of data.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *In SIGMOD Conference*, pages 1247–1250.

Craven, M. and Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In *ISMB-99*.

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L. S., and Weld, D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In Dekang Lin, et al., editors, *ACL*, pages 541–550. The Association for Computer Linguistics.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In Tony Jebara et al., editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196. JMLR Workshop and Conference Proceedings.

Marsh, E. and Sager, N. (1982). Analysis and processing of compact text. In *Proceedings of the 9th Conference on Computational Linguistics - Volume 1*, COLING '82, pages 201–206, Czechoslovakia. Academia Praha.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mintz, M., Bills, S., Snow, R., and 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*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peng, N., Poon, H., Quirk, C., Toutanova, K., and Yih, W.-t. (2017). Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Quirk, C. and Poon, H. (2016). Distant supervision for relation extraction beyond the sentence boundary. *CoRR*, abs/1609.04873.

Riedel, S., Yao, L., McCallum, A., and Marlin, B. M. (2013). Relation extraction with matrix factorization and universal schemas. In Lucy Vanderwende, et al., editors, *HLT-NAACL*, pages 74–84. The Association for Computational Linguistics.

Roller, R. and Stevenson, M., (2014). *Self-supervised Relation Extraction Using UMLS*, pages 116–127. Springer International Publishing, Cham.

Tsai, C.-T. and Roth, D. (2016). Concept grounding to multiple knowledge bases via indirect supervision. *TACL*, 4:141–154.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Number 8 in Structural analysis in the social sciences. Cambridge University Press, 1 edition.