Relation Extraction Using TBL with Distant Supervision

Maengsik Choi and Harksoo Kim

Program of Computer and Communications Engineering, College of IT,
Kangwon National University, Republic of Korea
{nlpmschoi, nlpdrkim}@kangwon.ac.kr

Abstract. Supervised machine learning methods have been widely used in relation extraction that finds the relation between two named entities in a sentence. However, their disadvantages are that constructing training data is a cost and time consuming job, and the machine learning system is dependent on the domain of the training data. To overcome these disadvantages, we construct a weakly labeled data set using distant supervision and propose a relation extraction system using a transformation-based learning (TBL) method. This model showed a high F1-measure (86.57%) for the test data collected using distant supervision but a low F1-measure (81.93%) for gold label, due to errors in the training data collected by the distant supervision method.

Keywords: Relation extraction, Transformation-based learning, Distant supervision

1 Introduction

In natural language documents, there are a huge number of relations between named entities. Automatic extraction of these relations from documents would be highly beneficial in the data analysis fields such as question answering and social network analysis. Previous studies on relation extraction [1,2,3,4] have been mainly conducted through supervised learning methods using Automatic Content Extraction Corpus (ACE Corpus) [5]. However, recent studies have investigated some methods based on simple rules rather than on complex algorithms because the supervised learning method using weakly labeled data generated by distant supervision has made it possible to use a large amount of data [6,7,8]. The distant supervision reduces the construction cost of training data by automatically generating a large amount of training data. In this paper, we propose a relation extraction system to generate weakly labeled data using DBpedia ontology [9] and classify the relations between two named entities by using transformation-based learning (TBL) [10].

2 Relation Extraction Method Based on TBL

As shown in Figure 1, the proposed system consists of three parts: (1) a distant supervision part that generates weakly labeled data from a relation knowledge base and

articles, (2) a training part that generates a TBL model by using the weakly labeled data as training data, and (3) an applying part that extracts relations from new articles using the TBL model.

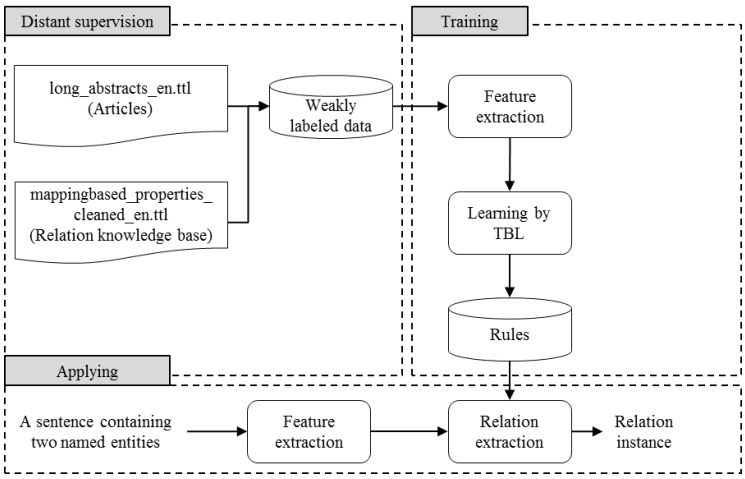

Fig. 1. Overall architecture of the proposed system

2.1 Construction of Weakly Labeled Data with Distant Supervision

For distant supervision, we use DBpedia ontology as a knowledge base. The DBpedia ontology is populated using a rule-based semi-automatic approach that relies on Wikipedia infoboxes, a set of subject-attribute-value triples that represents a summary of some unifying aspect that the Wikipedia articles share. As shown in Figure 2, the proposed system extracts sentences from Wikipedia articles by using the triple information of Wikipedia infobox.

Fig. 2. Example of an infobox and an article in Wikipedia

For example, the second sentence in Figure 2 is extracted from the weakly labeled data having a “Born” relation between “Madonna” and “Bay City, Michigan” and a “Residence” relation between “Madonna” and “New York City.”

2.2 Relation Extraction Using TBL

To extract features from the weakly labeled data, the proposed system performs natural language processing using Apache OpenNLP [10], as follows.

1. Sentences are separated using SentenceDetectorME.
2. Parts of speech (POS) are annotated using Tokenizer and POSTaggerME.
3. Parsing results are converted to dependency trees, and head words are extracted from the dependency trees.

For TBL, the proposed system uses two kinds of templates; a morpheme-level template and a syntax-level template. As shown in Figure 3, the morpheme-level template is constituted by the combination of words and POSs that are present in the left and right three words of the target named entities based on the POS tagging results.

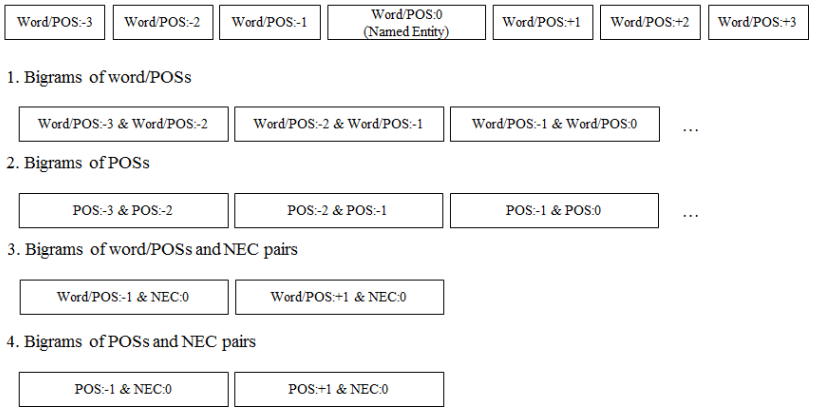

Fig. 3. Morpheme-level template

In Figure 3, the numbers described as "-n" and "+n" mean a left *n*-th word and right *n*-th word, respectively, from the target named entities. Then, "NEC" means the class name of the target named entity. As shown in Figure 4, the syntax-level template is constituted by the combination of two target named entities, common head word of the two named entities, head word (parent node in a dependency tree) of the two named entities, and dependent word (child node in a dependency tree) of the two named entities based on dependency parsing results.

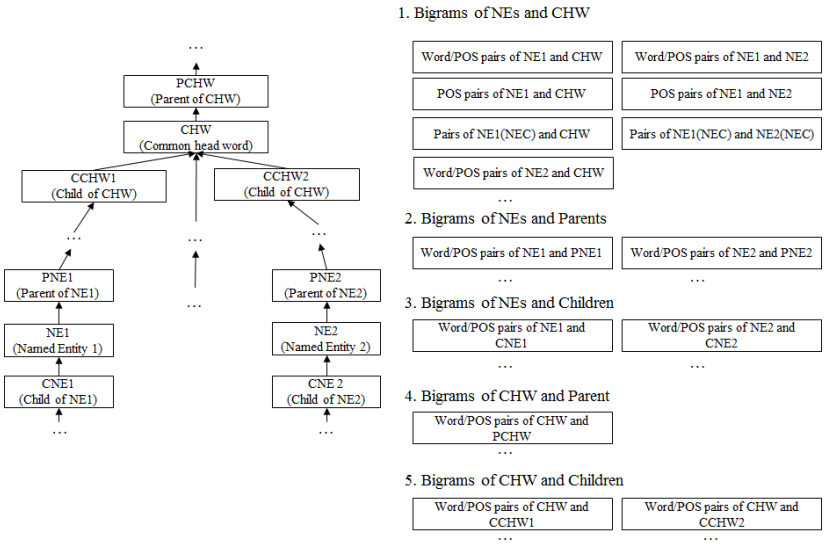

Fig. 4. Syntax-level template

3 Evaluation

3.1 Data Sets and Experimental Settings

We used DBpedia ontology 3.9 in order to generate weakly labeled data. Then, we selected the 'person' class as a relation extraction domain. Next, we selected five attributes highly occurred in the infobox templates of the 'person' class. Finally, we constructed a weakly labeled data set by using distant supervision. Table 1 shows the distribution of the weakly labeled data set.

Table 1. The number of instances in weakly labeled data

Attribute	All data	Weak-label test data	Gold-label test data
ActiveYearsStartYear	8,913	866	41
ActiveYearsEndYear	9,045	921	48
Award	1,627	140	4
BirthPlace	53,100	5,267	201
Nationality	8,754	869	133
Total	81,439	8,063	427

For the experiment, 90% of the weakly labeled data were used as training data, and the remaining 10% were used as test data (hereinafter "weak-label test data"). In addition, to measure the reliability of the weak-label test data, a gold-label test data set, which was manually annotated with correct answers after randomly selecting 822

items, was constructed. During manual annotation, 395 noise sentences out of 822 sentences were removed.

3.2 Experiment Results

As shown in Table 2, the proposed system showed high performance with regard to the weak-label test data but low performance with regard to the gold-label test data. This is because of noise sentences (*i.e.*, sentences that do not describe the relation between two named entities) that are included in the training data collected through the distant supervision method. For example, "Christopher Plummer" and "Academy Award" have an "Award" relation, and based on this, the extracted sentence ("The film also features an extensive supporting cast including Amanda Peet, Tim Blake Nelson, Alexander Siddig, Amr Waked and Christopher Plummer, as well as Academy Award winners Chris Cooper, William Hurt") describes the "Award" relation between "Chris Cooper and William Hurt" and "Academy Award." The results of analysis of the gold-label test data showed that there were 395 noise sentences out of 822 sentences.

Table 2. The performance of the proposed model

Data	Accuracy	Macro precision	Macro recall	F1-measure
Weak-label test data	0.9031	0.8735	0.8582	0.8657
Gold-label test data	0.7424	0.8275	0.8113	0.8193

The proposed system showed high performance for the weak-label test data. The fact reveals that the proposed system can show better performance for the gold-label test data only if good quality training data is ensured. In order to solve this problem, a method that can reduce the noise of the training data collected through distant supervision is required.

4 Conclusion

We proposed a relation extraction system using TBL based on distant supervision. In the proposed system, transformation rules were extracted based on a morpheme-level template that reflects the linguistic characteristics around named entities. They were also extracted based on a syntax-level template that reflects the syntactic dependency between two named entities. The experiment results indicated that higher performance could be expected only when the quality of the training data, which were extracted based on the distant supervision method, is improved. In the future, we will study on a method to increase the quality of training data collected by distant supervision.

5 Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2013R1A1A4A01005074), and was supported by ATC(Advanced Technology Center) Program "Development of Conversational Q&A Search Framework Based On Linked Data: Project No. 10048448". It was also supported by 2014 Research Grant from Kangwon National University(No. C1010876-01-01).

6 References

1. Culotta, Aron, and Jeffrey Sorensen: Dependency tree kernels for relation extraction. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. As-sociation for Computational Linguistics, (2004)
2. Bunescu, Razvan C., and Raymond J. Mooney: A shortest path dependency kernel for relation extraction. In: Proceedings of HLT/EMNLP, pp. 724-731 (2005)
3. Zhang, Min, Jie Zhang, and Jian Su. Exploring syntactic features for relation extraction using a convolution tree kernel. In: Proceedings of HLT-NAACL, pp. 288-295 (2006)
4. Zhou, G., Zhang, M., Ji, D. H., & Zhu, Q.: Tree kernel-based relation extraction with context-sensitive structured parse tree information. In: Proceedings of EMNLP-CoNLL pp. 728–736 (2007)
5. NIST 2007. The NIST ACE evaluation website. http://www.nist.gov/speech/tests/ace
6. Mintz, Mike, et al: Distant supervision for relation extraction without labeled data. In: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics, pp. 1003-1011 (2009)
7. Tseng, Yuen-Hsien, et al: Chinese Open Relation Extraction for Knowledge Acquisition. EACL 2014, pp. 12-16 (2014)
8. Chen, Yanping, Qinghua Zheng, and Wei Zhang.: Omni-word Feature and Soft Constraint for Chinese Relation Extraction. In: Proceedings of the 52nd Annual Meeting of the Associ-ation for Computational Linguistics, pages 572–581 (2014)
9. http://wiki.dbpedia.org/Downloads39
10. Ngai, Grace, and Radu Florian: Transformation-based learning in the fast lane. In: Proceed-ings of NAACL, pp. 40-47 (2001)
11. https://opennlp.apache.org/
12. Aprosio, Alessio Palmero, Claudio Giuliano, and Alberto Lavelli.: Extending the Coverage of DBpedia Properties using Distant Supervision over Wikipedia. In: NLP-DBPEDIA@ ISWC. (2013)
13. Choi, Maengsik, and Harksoo Kim.: Social relation extraction from texts using a support-vector-machine-based dependency trigram kernel. In: Information Processing & Management 49.1 pp. 303-311 (2013)