Proper Noun Learning from Unannotated Corpora for Information Extraction[1]

Seung-Shik Kang

School of Computer Science, Kookmin University & AITrc, Seoul 136-702, Korea
sskang@kookmin.ac.kr, http://nlp.kookmin.ac.kr/~sskang

Abstract. Named entity (NE) tagged corpus is an important resource for the learning of extraction patterns in information extraction system. We constructed Korean NE tagged corpus for economy, accident, and travel domain. As a semi-automatic approach to construct NE tagged corpus, a pattern learning method has been explored to extract personal names automatically from the raw corpus. Our NE tagging system has been trained for unannotated corpora to collect NE extraction patterns. Pattern extraction starts from a small set of proper names and NE extraction patterns are generated semi-automatically. Extracted patterns are used to automatically identify proper nouns from the text.

1 Introduction

Information retrieval system tends to extract too many search results for a given query. Manual selection of extracting useful information from search engines is a tedious work. Information extraction (IE) system automatically extracts predefined types of information from the extremely large set of information sources[1,2,3]. The target of discovering the knowledge in IE system is to find out meaningful named entities. Named entities are proper nouns that are extracted by IE system and they are the subjects of 5W1H. Typical named entities are person names, location names, organization names, product names, and numeric expressions like time, date, money, percent, and so on. So, the core function of the IE system is to identify the named entities and extract the predefined subject from the context of the text.

Information extraction started from the named entity contest of MUC and IREX workshop on Japanese text[4,5]. The major topics of the conferences are named entity extraction and coreference resolution on the specific domain. The basic language resources in IE system development are annotated corpora, language analyzer, and cue word dictionary. We constructed Korean NE tagged corpus on economy, accident, and travel domain. While constructing the corpus, we tried to automatically identify named entities from the text without using the lexicon or cue word dictionary. We trained our system on unannotated corpus to generate NE extraction templates, starting from the small set of proper names.

[1] This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Advanced Information Technology Research Center (AITrc).

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3036, pp. 670–674, 2004.

2 Corpus Construction and Annotation

We collected a raw corpus and defined a named entity tagset for the construction of named entity tagged corpus. Raw corpus of Korean articles has been collected from newspaper articles and webpages. They are 3,000 articles of news on economy, travel, social events, and seminars. Manual NE tagging is performed through (1) initially by automatic tagging system, (2) manual tagging of untagged entities using tagging tool, and (3) manual cross checking. NE tagset is based on the tagsets of MUC and IREX. Tagset classes of named entities are proper nouns and numeric expressions that are valuable for information extraction. Proper nouns are divided into specific classes and artifacts are divided into title and description. Numeric expressions are date, time, money, percent, and quantity. In addition, phone and address are added to numeric expressions. Named entity tags are (1) proper nouns(PERSON, LOCATION, ORGANIZATION, TITLE, DESCRIPTION, URL), (2) numeric expressions(DATE, TIME, MONEY, PERCENT, QUANTITY, PHONE, ADDRESS), (3) ambiguity, and (4) reference(referent and its antecedent relationship).

Constructing tagged corpus is a laborious work with high cost. For tagging efficiency, we developed NE/CO(Named Entity & Coreference) tagging tool. It provides a good tagging environment by marking a named entity block and selecting a tag in a graphic user interface. Fig. 1 shows an example of NE/CO tagging. The tagging tool provides a pre-tagging function that pre-tagged named entities are automatically tagged as a named entity. It minimizes the tagging errors and increases the accuracy. In addition, it has a statistical report generation function that generates a named entity list with its frequency counts. It also helps to find mis-tagged entities merely by checking the entities with low frequency.

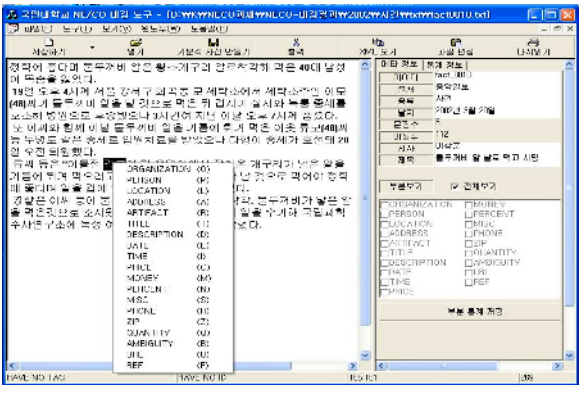

Fig. 1. NE/CO tagging workbench

3 Learning Extraction Patterns from Unannotated Corpora

There are two approaches in IE systems: knowledge engineering and automatically trainable approaches. Knowledge engineering technique is characterized as manually encoding extraction patterns. Knowledge engineers examine the corpus and write extract patterns. It requires lots of labors and the skill of a knowledge engineer affects on the performance of IE system. Automatically trainable approach needs an annotated corpus. Annotations are targeted to particular function like named entity recognition or coreference resolution. Once a suitable training corpus is constructed, the learning system automatically acquires extraction patterns from the corpus. In supervised learning system, the supervisor interacts with the learning system by indicating that the automatically acquired patterns are correct or not.

NE tagging system needs extraction patterns to identify named entities in the text. There are two tagging methods of manual construction and automatic construction. Manually constructed pattern is accurate and the performance of IE system is better than that of IE systems that are using automatically constructed patterns. Manual construction needs a continuous tuning with high cost. So, automatic construction is applied to get candidate patterns. Our NE tagging system has been trained for unannotated corpora to get NE extraction patterns that are included in the corpora. Extracting patterns start from a small set of proper names and NE extraction patterns are generated semi-automatically. Those extracted patterns are applied to automatically identify proper nouns when we construct NE tagged corpus.

The first step of the learning process of pattern templates is an identification of proper nouns that are given as an initial set of seeds. For each named entity that is given in the initial NE set, NE extraction patterns are automatically generated from the raw corpus. NE patterns are a sequence of part-of-speech or cue word. New patterns are mined from the corpus for the seed nouns in initial NE set, and they are added to the pattern set. So, extraction pattern set is expanded by repeating the learning process, and newly found proper nouns are added to the seed set. The pattern extraction process is repeated until no more patterns are added, or it stops after a predetermined number of iterations.

Pattern creation module adopts a model that is used in NE chunking/tagging and considers precedent or subsequent morphemes as contextual information. The template pattern consists only of a noun or a verb and others are not considered. When we generate a new pattern, named entity is marked as PERSON and the cue word is a morpheme. NE extraction of person names starts from the identification of named entities in the text that matches one of the extraction patterns. If a pattern is matched, then it is marked as a named entity. For the annotation of proper nouns, NE tagging system annotates proper nouns in the corpus by using NE extraction patterns. Named entity is described as a single word in the extraction pattern and multi-word proper nouns are combined into a noun phrase by NP chunking.

… Today	President	Kim … →	<null : President : PERSON>
adverb	noun	PERSON	extraction pattern

4 The Experiment

We performed an experiment for the automatic recognition of Korean person names. Test documents in this experiment are randomly selected from the raw corpus. Some person names that are found frequently in test corpus are given as an initial set of proper names. Experimentation has been performed for 2, 10, 20 person names as a seed name set, respectively. For the seed names, we run NE pattern extractor iteratively and the seed name set is expanded. The system runs until no more new names are added to the seed set. We automatically removed a pattern that has no information on both sides of the seed name and the patterns with common nouns. Table 1 shows the precision rates by the number of iterations. The precision ratio is 93%~94% for each experiment and the recall ratio is different according to the instances of the initial names. As a result, we found that NE pattern types depend on the document categories, and person names that are identified by NE patterns are also domain-dependent. When we set the initial seed set as politicians like 'Kim Dae-Jung' and 'Clinton', the system found most of the politically related person names, but it does not found the names of artists or novelist. Therefore, seed names should be given carefully that cover various categories of the documents.

Table 1. Precision of named entity extraction

no. seeds / iteration	5	10	20
1	0.935	0.939	0.938
2	0.965	0.938	0.947
3	0.967	0.926	0.946
4	0.967	N/A	N/A

5 Conclusion

We have constructed an NE tagged corpus for 3,000 articles of economy, accident, and travel domain. As a semi-automatic construction of the corpus, we applied automatic annotation of NE tags by using a proper noun learning technique of extracting person names from the raw corpus. NE extraction patterns are automatically collected from unannotated corpus, starting from a small set of seed names and expanding extraction patterns by iteration method. Our automatic tagging system will be extended to extracting common proper nouns regardless of the area.

References

1. Appelt, D. E. and David J. Israel, "Introduction to Information Extraction Technology", A Tutorial Prepared for IJCAI-99, 1999.
2. Cardie, C., "Empirical Methods in Information Extraction", AAAI-97, pp.65-79, 1997.

3. Riloff, E, “Information Extraction as a Stepping Stone toward Story Understanding”, In Computational Models of Reading and Understanding, Ashwin Ram and Kenneth Moorman, eds., MIT Press, 1999.
4. MUC, Proc. of 7th Message Understanding Conference(MUC-7), MUC, 1998.
5. Sekine, S., and Y. Eriguchi, “Japanese Named Entity Extraction Evaluation - Analysis of Results”, the 18th International Conference on Computational Linguistics (COLING'2000), pp.1106-1110, 2000.