Fine-Grained Named Entity Recognition and Relation Extraction for Question Answering

Changki Lee, Yi-Gyu Hwang, Myung-Gil Jang

Electronics and Telecommunications Research Institute (ETRI)

161 Gajeong-dong, Yuseong-gu, Daejeon, 305-350, Korea

{leeck, yghwang, mgjang}@etri.re.kr

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; Linguistic processing

General Terms: Algorithms, Experimentation.

Keywords: Question answering, fine-grained named entity recognition, relation extraction.

1. INTRODUCTION

In most QA systems, sentences or passages that are regarded as the most relevant to the question are extracted and then answers are retrieved by using NLP techniques (especially named entity recognition and syntactic pattern). Recent advances have brought some systems to within 90% accuracy when classifying named entities into broad categories, such as person, organization, and location. But more finely grained named entity recognition and relation extraction between these entities have additional advantages in QA [1, 2, 3, 7].

In most researches have gone into the coarse categorization of named entities and the syntactic relation pattern extraction, we are not aware of much previous work using machine learning algorithms to perform fine-grained NER and relation extraction between these fine-grained named entities.

Fleischman and Hovy describe a method for automatically classifying person instances into eight finer-grained subcategories [1]. They use a supervised learning method. But a training data is highly skewed, because they use a simple bootstrapping method to generate the training data automatically. And they classify only person and location, but we need to classify all kinds of NE types.

Mann explores the idea of a fine-grained proper noun ontology and its use in QA [2]. He builds a proper noun ontology from unrestricted text. The disadvantage of this method is that its coverage is small, because he uses simple textual co-occurrence patterns.

Agichtein and et al. show that a relatively small number of binary relationships account for the most of the queries in the sample [3]. But they define 12 relationships between coarse entities (9 classes) and do not show the relation extraction method.

Ravichandran and Hovy explore the surface text patterns for QA systems [7]. They use a pattern-learning algorithm with bootstrapping method. But they do not cope with various surface expressions of the sentence, because of the word ordering and long distance dependencies.

In this paper, we describe a fine-grained NER and relation extraction between fine-grained named entities and show the effect on QA.

2. FINE-GRAINED NER USING CRF

We define 147 fine-grained NE types in consideration of user's asking points for finding answer candidates of a QA system. They have 15 top levels and each top node consists of 2~4 layers. The base set of such types is; *person, study field, theory, artifacts, organization, location, civilization, data, time, quantity, event, animal, plant, material*, and *term*.

In this case, Conditional Random Fields (CRFs) can not be applied directly, because CRFs which have many classes are too time consuming. To solve this problem, we break down the NER task in two parts; boundary detection and NE classification.

Let $\mathbf{x} = <x_1, x_2,\ldots,x_T>$ be some observed input data sequences, such as a sequence of words of sentences in a document. Let $\mathbf{y} = <y_1,y_2,\ldots,y_T>$ be a set of NE states, each of which is associated with a NE type and a boundary (i.e. B and I). We define the conditional probability of a state sequence $\mathbf{y}$ given an input sequence $\mathbf{x}$ as follows:

$$P(\mathbf{y}\,|\,\mathbf{x}) = P(\mathbf{c},\mathbf{b}\,|\,\mathbf{x}) = P(\mathbf{b}|\mathbf{x})P(\mathbf{c}\,|\,\mathbf{b},\mathbf{x}) \approx P(\mathbf{b}|\mathbf{x})\prod_{i=1} P(c_i\,|\,b_i, x_i) \quad (1)$$

where $\mathbf{b} = <b_1, b_2, \ldots b_T>$ is a set of boundary states, $\mathbf{c} = <c_1, c_2, \ldots c_T>$ is a set of NE class states.

We define the conditional probability of a boundary state sequence $\mathbf{b}$ given an input sequence $\mathbf{x}$ using CRFs as follows:

$$P_{CRF}(\mathbf{b}\,|\,\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\exp\left(\sum_{t=1}^{T}\sum_k \lambda_k f_k(b_{t-1}, b_t, \mathbf{x})\right) \quad (2)$$

where $f_k(b_{t-1}, b_t, \mathbf{x})$ is an arbitrary feature function over its arguments, and λ_k is a learned weight for each feature function. Higher λ weights make their corresponding FSM translations more likely [4].

We calculate the conditional probability of a NE class c_i given a NE (b_i, x_i) extracted by boundary detector as follows:

$$P_{ME}(c_i\,|\,b_i,\mathbf{x}) = \frac{1}{Z(b_i,\mathbf{x})}\exp\left(\sum_k \lambda_k f_k(c_i, b_i, \mathbf{x})\right) \quad (3)$$

The primary advantage of CRFs over HMM is their conditional nature, resulting in the relaxation of independence assumptions required by HMMs. Additionally, CRFs avoid the label bias problem, a weakness exhibited by Maximum Entropy Markov Models (MEMMs). CRFs outperform both MEMMs and HMMs on a number of real-world sequence labeling tasks [4].

3. RELATION EXTRACTION

We define 37 binary relationships between fine-grained named entities in consideration of user's asking points. Table 1 shows the 10 most frequent relations observed in our training set.

Relation (z_{ij})	Subject (y_i)	Object (y_j)	Freq.
has_position	PS_NAME	CV_POSITION	4039
has_person	LCP_COUNTRY, OG_OTHERS,…	PS_NAME	3208
has_occupation	PS_NAME	CV_OCCUPATION	760
has_email	PS_NAME	TMI_EMAIL	524
has_business	LCP_COUNTRY, LCP_CITY,…	OGG_BUSINESS	351
has_building	LCP_COUNTRY, LCP_CITY,…	AF_BUILDING	251
has_county	LCP_PROVINCE, LCP_CITY,…	LCP_COUNTY	237
has_technology	LCP_COUNTRY, OGG_BUSINESS,…	TR_TECHNOLOGY	210
has_city	LCP_COUNTRY, LCP_PROVINCE	LCP_CITY	156
has_service	LCP_COUNTRY, OGG_BUSINESS,…	TMI_SERVICE	132

Table 1. Top 10 most frequent relations.

Let $\mathbf{z}= <z_{12},…,z_{ij},…>$ be a set of relation states, each of which is associated with two named entities (i.e. y_i and y_j). We define the conditional probability of two state sequences $\mathbf{z}$ and $\mathbf{y}$ given an input sequence $\mathbf{x}$ as follows:

$$P(\mathbf{z},\mathbf{y} \mid \mathbf{x}) = P(\mathbf{y}|\mathbf{x})P(\mathbf{z} \mid \mathbf{y},\mathbf{x}) \approx P(\mathbf{y}|\mathbf{x})\prod_{i,j} P(z_{ij} \mid y_i, y_j, \mathbf{x}) \quad (4)$$

where $P(\mathbf{y}|\mathbf{x})$ is the conditional probability of a NE sequence $\mathbf{y}$ given an input sequence $\mathbf{x}$ as defined by Equation 1.

We calculate the conditional probability of a relation state z_{ij} given two named entities (y_i, y_j) as follows:

$$P_{ME}(z_{ij} \mid y_i, y_i, \mathbf{x}) = \frac{1}{Z(y_i, y_i, \mathbf{x})}\exp\left(\sum_k \lambda_k f_k(z_{ij}, y_i, y_i, \mathbf{x})\right) \quad (5)$$

4. EXPERIMENT

The experiments for fine-grained NER were performed on our Korean fine-grained NE data set. The data set consists of 6,000 documents tagged by human annotators. We used 5,500 documents as a training set and 500 documents as a test set, respectively. We trained the model by L-BFGS using our C++ implementation of CRFs and Maximum Entropy (ME). We use a Gaussian prior of 1 (for ME) and 10 (for CRFs). We perform 500 iterations for training.

Task	Method	Training Time (second/iter.)	Pre (%)	Rec (%)	F1 (%)
Boundary detection	ME	**6.23**	87.4	80.6	83.9
	CRFs	24.14	**89.5**	79.7	**84.3**
NER	1 stage:ME(baseline)	154.97	82.5	73.5	77.7
	1 stage:CRFs	12248.33	-	-	-
	2 stages:ME+ME	**24.59**	82.8	73.3	77.8
	2 stages:CRFs+ME	42.50	**83.2**	**74.5**	**78.6**

Table 2. Performance of fine-grained NER (147 classes).

Table 2 shows the performance of boundary detection and fine-grained NER using CRFs and ME. In boundary detection, we obtained the performance, 83.9% F1 and 84.3% F1 using ME and CRFs respectively. In NER, our proposed model (2 stages: CRFs+ME) obtained 78.6% F1. The baseline model (1 stage: ME)

obtained 77.7% F1. We also reduced the training time to 27% without loss of performance compared to the baseline model.

The experiment for relation extraction was performed on our Korean relation extraction data set which consists of 485 documents tagged by human annotators. We used 455 documents (one-year IT news articles) as a training set and 30 documents (one-month IT news articles) as a test set, respectively.

Relation	Precsion	Recall	F1
has_position	86.2	73.5	79.4
has_person	88.2	85.7	87.0
has_occupation	87.0	83.3	85.1
has_email	65.2	60.0	62.5
has_business	55.6	83.3	66.7
Overall	**80.8**	**66.2**	**72.8**

Table 3. Performance of Relation Extraction.

Table 3 shows the performance of the relation extraction for the 5 most frequent relations and the overall performance of 37 relations.

To show the effect of the fine-grained NER and relation extraction, we performed another experiment for a QA system. We modified ETRI QA Test Set [5], which consists of 402 pairs of question and answer in encyclopedia, as the IT news domain and used 50 factoid questions in the experiment.

QA system	MRAR
Passage retrieval + Coarse NER (9 classes)	0.538
Passage retrieval + Fine-grained NER (147 classes)	0.660
Passage retrieval + Fine-grained NER + Relation extraction	0.716

Table 4. Performance of Question Answering.

Table 4 shows the performance of QA systems. The QA system with fine-grained NER and relation extraction achieved about 48% improvement over QA with coarse NER.

5. CONCLUSION

In this paper, we describe a fine-grained NER and relation extraction using CRFs and ME for QA. Using the proposed approach, we obtained 78.6% F1 for 147 fined-grained NE types and 72.8% F1 for 37 relations. In the QA system, the performance with fined-grained NER and relation extraction archived about 33% improvement over QA with coarse NER.

6. REFERENCES

[1] M. Fleischman and E. Hovy. *Fine grained classification of named entities*. COLING, 2002.

[2] G. Mann, *Fine-Grained Proper Noun Ontologies for Question Answering*, SemaNet'02: Building and Using Semantic Networks, 2002

[3] E. Agichtein, S. Cucerzan, and E. Brill. *Analysis of Factoid Questions for Effective Relation Extraction*, SIGIR, 2005.

[4] S. Fei, F. Pereira. *Shallow Parsing with Conditional Random Fields*, HLT & NAACL, 2003.

[5] H. Kim, J. Wang, C. Lee, C. Lee, M. Jang. *A LF based Answer Indexing Method for Encyclopedia Question Answering System*, AIRS, 2004.

[6] K. Han, H. Chung, S. Kim, Y. Song, J. Lee, H. Rim. *Korea University Question Answering System at TREC 2004*, TREC, 2004

[7] D. Ravichandran and E. Hovy. *Learning surface text patterns for a question answering system*, ACL, 2002