Named Entity Recognition using Machine Learning Methods and Pattern-Selection Rules

Choong-Nyoung Seon, Youngjoong Ko, Jeong-Seok Kim†, and Jungyun Seo
Department of Computer Science, Sogang University
1 Sinsu-dong, Mapo-gu, Seoul, 121-742, Korea
†Department of English Education, Yeungnam University
Kyoungsan-si, Kyoungsangbuk-do, 712-749, Korea
{wilowisp, kyj}@nlprep.sogang.ac.kr, uconnkim@yu.ac.kr, seojy@ccs.sogang.ac.kr

Abstract

Named Entity recognition, as a task of providing important semantic information, is a critical first step in Information Extraction and Question-Answering system. This paper proposes a hybrid method of the named entity recognition which combines maximum entropy model, neural network, and pattern-selection rules. The maximum entropy model is used for the proper treatment of unknown words, and neural network for disambiguation. The pattern-selection rules are used for the target word selection and for grouping of adjacent words. We use the data only from a training corpus and a domain-independent named entity dictionary so that our system, it is predicted, is applicable in any other domain. In addition, since each module of our system is independent, a new method can be easily adopted for executing each module.

1 Introduction

Named Entity (NE) recognition is a task in which person names, location names, organization names, monetary amounts, time, and percentage expressions are recognized in a text document. This task is a basic and important technique for Information Extraction (IE) and Question-Answering System.

Time, monetary amounts, and percentage expressions are fairly predictable. Hence, they can be processed most efficiently with finite state methods (Roche E., et al.,1997). But person names, location names, and organization names are highly variable because they are open classes. Still worse, it is much more difficult to recognize them because of unknown words and ambiguity problems.

The ambiguity problem between location names and organization names has drawn a particular attention in Korean. Let us illustrate this point:

Example 1: the Blue House as the Korean government

"cheng-wa-day *ey-se say nay-kak ul*
(Korean government) (PP:from) (new) (cabinet) (PP)
bal-phyo-hay-ta"
(announced)

("**The Blue House** announced the new cabinet")

Example 2 : the Blue House as the Korean President mansion

"tay-thong-lyeng *un* ***cheng-wa-day***
(the President) (PP) **(Korean President mansion)**
ey-se ches ep-mwu lul si-cak-hay-ta"
(PP:from) (first) (business) (PP) (began)

("The President began his first business in **the Blue House**")

In the first example, "*cheng-wa-tay* (the Blue House)" is tagged as an organization name, meaning the Korean government. In the second, it is a location name, meaning the Korean President mansion. To disambiguate the meaning of "*cheng-wa-tay* (the Blue House)", complex information such as contextual or lexical information is required. Still worse, there are many cases which even Korean native speakers cannot disambiguate, and to which they cannot assign proper tags.

Recent researches have been focused on improving the accuracy of NE recognition with several different techniques. Among others, there are Maximum Entropy Models (MEM) (Borthwick et al., 1998), Hidden Markov Models (HMM) (Bikel et al,. 1997), Decision Tree Model (Sekine et al., 1998), rule-based systems (Aberdeen et al., 1995; Krupka et al., 1998; Kyung Hee Lee et al., 2000), and hybrid systems (Srihari et al., 2000).

A system based on handcrafted rules may provide the best performance. But such a system requires painstaking skilled labor, and the rules have to be changed according to each application domain. HMM is generally regarded as the most successful statistical modeling method, but it requires a large size of corpus. Since learning methods like MEM and neural network can deal with the data sparseness problem effectively, a high accuracy can be achieved by using these methods without a large amount of corpus.

In this paper, we propose a hybrid method of maximum entropy model, neural network and pattern-selection rules in order to recognize the Korean NE. In section 2, we describe the structure of the proposed system and each moudule in the proposed system. Section 3 is devoted to the discussion of experiment results. In section 4, conclusion and future works are presented.

2 Named Entity Recognition System

The proposed system consists of five modules as shown in Figure 1.

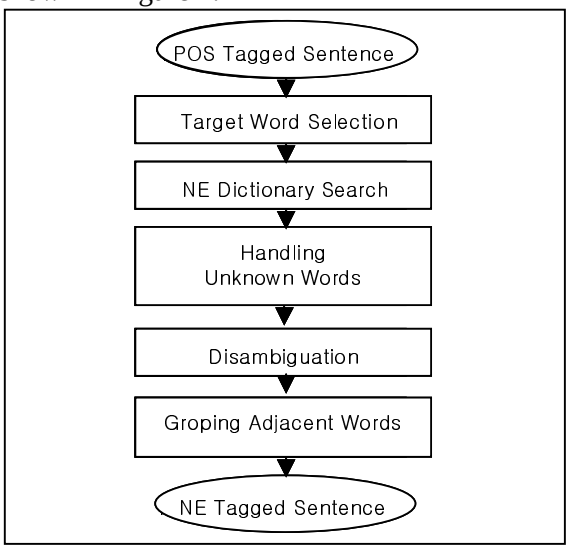

Figure 1 : Structure of the proposed system

The first module selects target words using Korean POS tags and clue word dictionary. The second module searches for target words in the NE dictinary. Then the third module handles unknown words using the MEM method with lexical sub-pattern information and a clue word dictionary. The second and third modules assign each target word to a NE tag or tentative duplicate tags (four type tags: person/location tag, location/organization tag, person/organization tag, and person/location/organization tag). The next module solves the ambiguity probelm using neural network. The features used in neural network are selected from the adjacent POS tags and the clue word dictionary. Finally, the last module converts adjacent words into a NE tag using pattern-selection rules.

This research aims to recognize only NE tags which are limited to person names, location names, and organization names: These three NE names are significant categories of MUC (Message Understanding Conference)-standard NE tags. It is straightforward that finite state methods can recognize the other NE tags. However, for a real information extraction system, the above three NE tags may not be enough. Thus, we pre-defined sub-categories for person names, location names, and organization names as follows:

Table 1 : Pre-defined sub-categories

Category	Sub-categories
Person	academic person, economic person, military person, religious person, political person, professional person, relational person, others
Location	country, state, city, province, continent, lake, river, mountain, geographic location, sight-seeing place, building, others
Organization	country, state, city, company, political organization, school, laboratory, association, department, mass media, others

NE tags related to these sub-categories are assigned to a target word only by *the NE dictionary search* module and *the grouping adjacent words* module.

2.1 Selecting Target Words for NE

The first letter of proper nouns in English are upper characters. Thus, we can easily find target words for NE. However, in Korean, (proper) nouns do not have the distinction of upper/lower characters. Still worse, Korean compound nouns are highly productive. Therefore, it is not a simple procedure to select target words for NE in Korean.

In Korean, the candidates for a target word are proper nouns, English characters and compound nouns. But the compound nouns with any proper noun are excluded from the candidates because they are handled in the Grouping Adjacent Words module.

To find target words, we construct a Trie dictionary. It is composed of the sequence of POS tags and the information of clue words. We suppose that the compound nouns for target words must have a clue word in the last common noun. Therefore, we can select target words when any pattern of compound nouns, proper nouns and English characters are found in input sentences. For example, "*Nong-uh-chon* (farming and fishing villages) [common noun] *jinhung* (promotion) [common noun] *kong-sa* (a public corporation) [common noun, *organization clue word*]" makes an entry (common noun : common noun : common noun-*organization clue word*) in the Trie dictionary.

2.2 Searching for target words in the NE Dictionary

The NE dictionary consists of a general NE dictionary and a domain NE dictionary. The general NE dictionary is constructed manually and the domain NE dictionary from train corpus automatically. The general NE dictionary is composed of three categories (person, location, and organization). Among these three categories, the location and organization categories share the same sub-categories enumerated in Table 1. But the person category is composed of only full name, first name and last name sub-categories (cf. Table 1). The full names were collected from "Seoul Telephone Directory", and the first names and the last names were automatically extracted from those. The location and organization names were collected from various web pages (e.g. Yahoo Weather Center) and books (e.g. Middle and High school geography book). Table 2 shows the size of the NE dictionary.

Table 2 : Size of the NE dictionary

	The number of entities (General)	The number of entities (Domain)
Person	422,151	278
Location	44,324	243
Organization	64,633	254

The target words, extracted in the target word selection module described in section 2.1, are looked up in the NE dictionary. When a target word is found in only one sub-category of the NE dictionary, it is tagged as the sub-category. If a target word is found from two or more sub-categories which belong to the different categories, it has a duplicate tag: We suppose that there is no ambiguity among the sub-categories in the same category. The ambiguity of the target words will be resolved by the disambiguation module using neural network.

2.3 Handling Unknown Word

The proper nouns like person names, location names, and organization names form an open set because they are created continuously. Therefore, they arise an out-of-entry word problem, which we call the '*unknown word problem*'.

In order to solve this problem, we use MEM, which is a powerful tool used in the situation where several ambiguous information sources need to be combined. There are two types of feature function template. One type uses lexical sub-patterns extracted by the NE dictionary and the other type clue words after target word.

In Korean, there are many lexical sub-patterns from Chinese characters which belong to ideography. Therefore, they are likely to be clues in many cases. We extract these lexical sub-patterns from the entries of the NE dictionary discussed in section 2.2. We restrict the number of candidate syllables to two from the first syllable and two from the last syllable of a unknown word. As an example of the clue lexical sub-patterns with the first two syllables, a lexical sub-pattern "*nam-bwu~* (the South)" of "*nam-bwu-the-mi-nel* (the South terminal)" is a clue lexical sub-pattern indicating a location name. As an example with the last one or two syllables, a lexical sub-pattern "*~si* (city)" of

"*se-wul-si* (Seoul city) is a clue lexical sub-pattern indicating a location name, and "*~hak-kyo* (school)" of "se-kang-tay-hak-kyo (the Sogang university)" is indicating an organization name. To select the clue lexical sub-patterns, we simply measure their validity as a feature of each NE category, using the difference of frequency between a NE category and the other categories. Then the extracted candidate syllables are sorted according to the decreasing order of their validity. We use only the syllables with validity value above the proper threshold value as clue lexical sub-patterns.

The feature function templates using lexical sub-patterns are shown in formulae (1) and (2).

$$f(history, tag) = \begin{cases} 1 \ if \ WORD = \_, PLOFLAG = \_, \\ \quad and \ \ tag = \_ \\ 0 \qquad\qquad\qquad\quad \text{Otherwise} \end{cases} \quad (1)$$

$$f(history, tag) = \begin{cases} 1 \ if \ PLOFLAG = \_, and \ \ tag = \_ \\ 0 \qquad\qquad\quad \text{Otherwise} \end{cases} \quad (2)$$

In the above formulae, *"WORD"* denotes a clue lexical sub-pattern. *"PLOFLAG"* is a flag, representing that the clue lexical sub-pattern belongs to any NE category. Here "tag" represents one of the three possible tags (person, location, and organization). The symbol "_" denotes any possible values.

In many cases, clue words are adjacent to a NE in a sentence. Thus we also constructed a clue word dictionary, as shown in Table 3. We extracted the clue words of each category from the various web pages (e.g. government web pages for political person name). Also, we used newspaper articles and other corpus to extract the clue words.

If a word with the POS tag of common noun or suffix is located after a target word, it is looked up in the clue word dictionary. The result is used as a feature in feature function template as shown in the following formula (3).

Table 3 : Clue word dictionary

Category	# of entities	Examples
Academic person	25	*kyo-swu* (professor), *sen-sayng-nim* (teacher)
Economic person	52	CEO, CTO, *koa-cang* (director)
Relational person	143	*a-pe-ci* (father)
Political person	486	tay-thong-lyeng (the President)
Military person	24	*so-day-cang* (platoon leader)
Religious person	14	*mok-sa* (clergyman)
Professional person	95	*ti-ca-i-ne* (designer)
Country	12	*kong-hoa-kwuk* (republic)
City	3	*swu-to* (capital)
State	2	*to* (state)
Administrative district	6	*ka*, *kwun*
Area	3	*ci-pang* (district)
Sight-seeing place	25	CC, *kong-wuen* (park)
Geographic location	41	*kang* (river), *san* (mountain)
Building	6	*pil-ding* (building), *man-syen* (mansion)
Association	10	*yen-hap* (union)
Company	127	*ken-sel* (construction)
Laboratory	5	*yen-kwu-sil* (laboratory)
Mass Media	30	TV
School	14	*tay-hak-kyo* (university)
Political organization	22	*kem-chal-cheng* (the public prosecutors office)

A feature function template using a clue word is as follows:

$$f(history, tag) = \begin{cases} 1 \ if \ CLUE = \_, and \ \ tag = \_ \\ 0 \qquad\qquad\quad \text{Otherwise} \end{cases} \quad (3)$$

In formula (3), *"CLUE"* represents a kind of categories in the clue word dictionary.

A maximum entropy solution for probability has the following form (Rosenfeld,1994; Ratnaparkhi,1998):

$$p(tag \mid history) = \frac{p(tag, history)}{\sum_{tag} p(tag, history)} \quad (4)$$

$$P(tag, history) = \frac{\prod_i \alpha_i^{f_i(history, tag)}}{Z(history)} \quad (5)$$

$$where \quad Z(history) = \sum_{tag} \prod_i \alpha_i^{f_i(history, tag)}$$

Any target word can have one of three tags only when the result value is more than a pre-set threshold value. In addition, if the difference between the maximum value and the second high value is less than a pre-set threshold value, the target word in this case will have a duplicate tag. These threshold values are decided empirically.

2.4 Resolving the disambiguation of the NE with a duplicate tag

In the above two sections, we have seen that the target words with the ambiguity have a duplicate tag. The duplicate tag is composed of four types; person/location tag, location/organization tag, organization/person tag, and person/location-/organization tag. Therefore, we learned the four kinds of neural network for each case and used them for solving the ambiguity problem.

We used the *SNNS 4.2* for neural network tool and the standard Backpropagation algorithm for the learning algorithm (*SNNS User Manual 4.2*). The structure of each neural network consists of input layer with 81 neurons, hidden layer with 42 neurons and output layer with 2 or 3 neurons (3 neurons for only a duplicate tag among 3 categories).

The input patterns of each network consist of two parts. One part uses POS tag information, and the other part uses lexical information.

The POS tag information adjacent to a target word is considered as significant features. After we remove useless POS tags like adverb, we extract POS tag information within the scope of the two POS tags on the left and the two POS tags on the right of the target word (Uchimoto et al., 2000). Then we define the useful tag sets in each position and uses them as input features. The total number of input features using POS tag information is 55.

We also extract the lexical information with the same scope except verb lexical information. For this purpose, we use a new clue word dictionary with additional five categories which is an extended version of the clue word dictionary in Table 3. Finally, a total of 26 features represents whether a given word belongs to the clue word dictionary. Table 4 lists the added categories of the new clue word dictionary.

Table 4: Added clue word dictionary

Category	# of entities	Examples
Person clue	28	*sin-im* (new appointment), *ui-wuen* (member)
Location clue	77	*ma-ul* (village), *twul-lay* (around)
Organization clue	52	*kwuk-pep* (national law), *tan-chay* (group)
Location verb clue	46	*tte-na-ta* (leave), *to-chak-ha-ta* (arrive)
Organization verb clue	82	*Pal-phyo-ha-ta* (announce), *kay-choi-ha-ta* (hold)

Since the entities of the person, location, and organization clue categories in Table 4 does not have a proper meaning corresponding to any category in Table 3, they cannot have any category in Table 3. However, since we regarded these entities as the important clue words for disambiguation, we constructed these three clue categories. The location and organization verb clue categories are mainly used for resolving the ambiguities between location names and organization names.

All feature values used in neural network are binary.

2.5 Grouping Adjacent Words into a NE tag by Pattern-selection rules

Through the above disambiguation module, we can decide a NE tag within one word. But, in some case like "*kim-day-cwung* (Kim Dae-jung) *tay-thong-lyeng* (the President)", the meaning can become more clear when "*kim-day-cwung* (Kim Dae-jung)" is combined with the adjacent clue word, that is, "*day-thong-lyeng* (the President)". Finally, a word in this case can be tagged into a detailed NE sub-category through this module.

To group the adjacent clue words into one NE tag, we automatically extract pattern- selection

rules from training corpus. To extract pattern-selection rules, we use the NE tag information, the lexical information, the clue word dictionary in Table 3, and the POS tag information. Finally, we obtain a total of 191 pattern- selection rules.

A sample pattern-selection rule is shown as follows:

[Political person] = [Person] + {political CLUE}

Example : <*kim-day-cwung* (Kim Dae-jung) [Person] *tay-thong-lyeng* (the President) [CLUE:Political person]> **[Political person]**

3 Evaluation of experiment

3.1 Experiment settings

We used the KAIST (Korea Advanced Institute of Science and Technology) tagged corpus, which consists of two kinds. One (Corpus 1) is made of newspaper editorials, and the other (Corpus 2) is selected from novels. Therefore, we could evaluate our system in two different application domains. Table 5 and 6 show the settings of experiment data in details.

Table 5: Setting experiment data

	Corpus 1		Corpus 2	
	# of sentence	# of NEs	# of sentence	# of Nes
Train	2,555	1,471	6,108	1,678
Test	412	263	999	236

Table 6 : The number of each NE in corpus

	Corpus 1			Corpus 2		
	P	L	O	P	L	O
Train	337	133	1001	677	591	410
Test	26	40	197	102	69	65

where “P” indicates Person name, “L” Location name, and “O” Organization name.

3.2 Experiment results

We evaluated our system according to each corpus. The results are shown in Table 7. The target word with a duplicate tag may be regarded as the correct response in a case where any possible two or three NE tags of its duplicate tag become a correct response. We define the highest numerical value of the case as the *maximum recall*. More precisely, the maximum recall value represents the highest recall value obtained at the current module.

Table 7: Results of NEs recognition

		Corpus 1	Corpus 2	1+2
NE Dictionary Search	p	97.80%	96.83%	97.24%
	r	33.84%	51.69%	42.28%
	mr	88.97%	93.64%	91.18%
Unknown handling	p	93.64%	96.12%	94.98%
	r	39.16%	52.54%	45.49%
	mr	96.58%	94.49%	95.59%
Disambi-guation	p	83.77%	81.30%	82.63%
	r	84.41%	79.24%	81.96%
	mr	84.41%	79.24%	81.96%
	F	84.09%	80.27%	82.3%

where “p” denotes precision, “r” recall, “F” F-measure, and “mr” maximum recall.

We did not tune our system to each corpus. The comparison of the experiment results showed that the performance of Corpus 1 was nearly three points better than that of Corpus 2. Therefore, we found that the performance in the specific domain like editorials is better.

Table 8 shows the results of each NE category (Person, Location and Organization).

Table 8: Results in each NE category

	Person	Location	Organization
precision	91.04%	71.08%	82.01%
recall	95.31%	54.13%	87.02%

The performance of location names is the lowest.

The results of the disambiguation module are shown in Table 9.

Table 9: Results for disambiguation

	Corpus 1	Corpus 2
	Precision	Precision
Person	66.67% (2/3)	00.00% (0/4)
Location	38.89% (7/18)	65.71% (23/35)
Organization	82.09% (110/134)	64.52% (40/62)
Total	76.77% (119/155)	62.38% (63/101)

When we comparing with the results of each domain, we obtain the similar performance.

Table 10: Results of grouping adjacent words

Precision	Recall
90.47%	64.41%

Table 10 lists the results of grouping adjacent words module. Since we automatically selected pattern-selection rules only from training corpus, recall showed a lower performance in comparison with precision. However, this lower performance of recall does not necessarily threaten the validity of our research. That is, precision is more significant in that the aim of grouping adjacent words module is to add detailed tag information (sub-category).

4 Conclusion

This paper has discussed the recognition of named entities on the basis of a maximum entropy model, a neural network, and pattern-selection rules. The first step of the proposed method includes a target word selection module and NE dictinary search module. Then our method excutes a process for handling unknown words using MEM. In the next step, it solves a ambiguity problem using neural network. Finally, adjacent words are combined into one NE tag using pattern-selection rules. These pattern-selection rules are automatically acquired from a training corpus and a domain-independent NE dictionary.

All data, used in our system, are extracted only from a tagged training corpus and a domain-independent NE dictionary. Therefore, our system can be easily shifted into any other application domain without any significant effort and performance degradation. In addition, our system consists of independent modules. Thus, we expect a new method to be easily applied to each module.

The experiment result shows that an F-measure of 84.09% is achieved for the specific domain (Corpus 1: editorials), and an F-measure of 80.27% for the general domain (Corpus 2: novels). We found that the better performance is achieved in the editorial domain.

There are several possible future researches. First, since we extract all data from the training corpus and NE dictionary, we should collect and revise more tagged corpus and NE dictionary. Next, we should study more effective features for the maximum entropy model and the neural network model.

Acknowledgments

This work was supported in part by the Brain Korea 21 project sponsored by the Korea Research Foundation.

References

Aberdeen J., Burger J., Day D., Hirschman L., Robinson P. and Vilain M., 1995, MITRE: Description of the Alembic system used for MUC-6. In *Proceedings of 6th Message Understanding Conference (MUC-6)*, pp. 141-155.

Bikel D.M., 1997, Nymble: a high-performance learning name-finder, In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp.194-201, Morgan Kaufmann Publishers.

Bothwick A., et al., 1998, Description of the MENE named Entity System, In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Krupka G.R.and Hausman K., 1998, IsoQuest Inc: Description of the NetOwl Text Extraction System as used for MUC-7 In *Proceedings of Seventh Message Understanding conference (MUC-7)*.

Kyung Hee Lee, et al., 2000, Study on Named Entity Recognition in Korean Text, In *Proceedings of the 13th National Conference on Korean Information Processing*

Ratnaparkhi A., 1998, Maximum Entropy Models for Natural Language Ambiguity resolution, *PHD thesis*, Univ. of Pennsylvania.

Roche E. and Schabes Y., 1997, Finite-State Language Processing, *The MIT Press, Cambridge, MA.*

Rosenfeld R., 1994, Adaptive Statistical language Modeling, *PHD thesis*, Carnegie Mellon University.

SNNS User Manual, Version 4.2

Sekine S., Grishman R., and Shinnou H., A decision tree method for finding and classifying names in Japanese texts. In *Proceedings of 6th Workshop on Vary Large Corpora*, 1998.

Srihari R. Niu, C. and Li W., 2000, A Hybrid Approach for Named Entity and Sub-Type Tagging, In *Proceedings of 6th Conference on Applied Natural Language Processing (ANLP)*, pp. 247-254.

Uchmoto K., Ma Q., Murata M., Oasku H. and Isahara H., 2000, In *Proceedings of 38th Annual Meeting of the Association for Computational Linguistics*, pp. 326-335.