HMM based Korean Named Entity Recognition

Yi-Gyu Hwang, Eui-Sok Chung, Soo-jong Lim

Speech/Language Technology Research Center
Electronics and Telecommunications Research Institute
161, Gajeong-Dong, Yuseong-Gu, Daejeon, 305-350, Korea
{yghwang, eschung, lsj}@etri.re.kr

ABSTRACT

In this paper, we present a named entity recognition model for Korean Language. Named entity recognition is an essential and important process of Question Answering and Information Extraction system. This paper proposes a HMM based named entity recognition using compound word construction principles. In Korean, above 60% of NE (Named-Entity) is a compound word. This compound word may be consisted of proper noun, common noun, or bound noun, etc. There is an inter-contextual relationship among nouns which consists NE. NE and surrounding words of NE have a contextual relationship. For considering these relationships, we classified nouns into 4 word classes (Independent Entity, Constituent Entity, Adjacent Entity, Not an Entity). With this classification, our system gets contextual and lexical information by stochastic based machine leaning method from a NE labeled training data. Experimental result shows that this approach is better approach than rule-based in the Korean named-entity recognition.

Keywords: Named Entity Recognition, Compound Word Construction, HMM, Independent Entity, Constituent Entity, Adjacent Entity

1. Introduction

As the amount of information has been increased exponentially, it is very important task to find words like proper nouns, time and money expressions in the documents. Such needs give rise to **Named Entity Recognition (NER)**. Named entity recognition is a classification and identification process of person, location, and organization name (PLO) or numerical expressions.

However, there are two main problems in NER. First, Named Entity (NE) is included in an open word class. Person, location, and organization names can be newly made by the human. Adding such words to dictionary is a very time consuming task and it is impossible to all NE to add a dictionary. Second, NE can be used as several NE types (ambiguity problem). For example, "Paris" may be used as a **person** in the sentence "Paris was a prince of Troy." and used as a **location** in the following sentence "Paris is a capital city of France."

In the Korean NER, we must consider two main issues. First, the notable features of Korean is a few of orthographic property compared to English or Japanese. This causes difficulty in applying statistical methods in Korean NER. Second, there are many inter-word relationships and contextual relationships in the NE.

There are two main approaches in the NER such as a stochastic based and a rule based approach[3,

9]. In the stochastic criteria, there are Hidden Markov Model[1], Maximum Entropy Model[2, 8], Decision Tree/List Model[5, 6], and Hybrid Model[7].

Most of English NER systems used a statistics based approach and Korean systems used rules with a NE dictionary. Rule based NER system can achieve the proper performance with ease. But, in this system, rules are made by handcrafted. So, it is domain-specific and less portable. This means the rules must be modified according to the domain.

We investigated the structure of Korean NE and found combination principles of the NE. According to these principles, we classified nouns into 4 classes and recognized Korean NE. Our system included these principles in the Hidden Markov Model. In section 2, we describe the characteristics of Korean NE. Section 3 explains proposed model and section 4 shows an experimental result. Finally, conclusion and future works are presented in the section 5.

2. Characteristics of Korean Named Entity

2.1 Distributions of NE in Korean Documents

In Korean, several words are combined into and form a NE. Especially, location and organization names can be decomposed into many words. We analyzed 300 documents that are made up of economy articles (DS1), public performance articles (DS2), and web pages of trip guide (DS3). Table 1 shows the distributions of the NE. And, we found that there are about 66% of NEs that can be decomposed into more than one morpheme and 14.4% of NEs more than 4 morphemes. Here is a Table 2 which shows the length of NEs in Korean.

Table 1. NE distributions in the documents

NE type	DS1	DS2	DS3
Person	8.6%	28.8%	2.8%
Location	17.1%	15.5%	61.2%
Organization	24.6%	11.3%	4.0%
Date	20.2%	18.3%	5.2%
Time	0.7%	6.3%	1.6%
Price	5.5%	1.9%	3.8%
Percent	11.7%	0.2%	0.1%
Quantity	11.7%	14.7%	17.6%
Phone	0.00%	3.1%	3.7%

Table 2. Length of NEs

Morpheme leng. of NE	percent
n = 1	33.1%
n = 2	40.0%
n = 3	12.5%
n = 4	7.4%
n = 5	4.4%
n ≥ 6	2.6%

Also, we found that the part-of-speech (POS) distributions of the NE's boundaries are as follows and detailed statistics will be examined further in the next section.

- There are 32.1% and 36% of nouns in the front and back of NEs respectively.
- 55.1% and 49.4% of them was directly related to the NER.

2.2 Classification of Nouns for NE Recognition

Above investigation means surrounding nouns in the NEs are very important clues to recognize a NE. Standing on this analysis result, we divided nouns into 4 groups by its role in the NER.

- **Independent entity** (IE), a noun that can be solely a NE.
  - Ex1) Paris, Bush, Pentagon, Motorola, …
  - Ex2) *"Kim-dai-cung"*, *"Seoul"*, …

- **Constituent entity** (CE), a noun that can not be solely a NE but can be combined with other nouns and forms a NE. For example, the company suffix is one of the CE class.
  - Ex1) Co., Soft, Electronics, …

- Ex2) *"kong-sa"* (public company), *"kang"* (river), *"kong-hang"* (airport), *"keuk-cang"* (theater), …

- **Adjacent entity** (AE), a noun that can be occurred in front or back of the NE and indicates following or front words sequences can be a NE. "Mr." and position titles are included in the AE class.
  - Ex1) Mr., CEO, vocalist, governor, district, …
  - Ex2) *"dan-won"* (member), *"sa-cang"* (CEO), *"chul-cin"* (origin), …

- **No entity** (NoE), a noun that is not an IE, CE or CE. Most of common nouns are included in this class.

2.3 Structural Regularity in NE

Most of NE has a structural regularity. Following combinations are possible structures of NE.

- IE, CE, and AE composed together and became an NE.
- Common noun or numeral noun is combined with SE[1].
- Unknown words are combined with IE, CE, or AE.

In the rule-based system, above structures are represented by rules which are manually made. Table 3 is the structures of NE which are derived from training data.

Table 3. Inter-word structures of NE

NE constituent type	percent
nn ce_date	5.5%
nc nc	3.9%
nc ce_location	3.6%
nn nc	2.6%
nn nb	2.4%
nn ce_percent	2.3%
nn ce_quantity	2.0%
ie_location ce_location	1.7%
nn s nn s nn	1.2%
nn ce_money	1.1%
nn ce_money ce_money	1.1%
nn ce_time	1.0%
mm nc	1.0%
ie_location nc	1.0%
ie_time nn ce_time	0.9%
nn s nn	0.7%
ie_time nn ce_time nn ce_time	0.7%
nc nc nc	0.7%
nn ce_quantity nc	0.6%
ie_location ie_location	0.5%
nn ce_date nn ce_date	0.5%
ie_location nc nc	0.5%
ie person ie person	0.5%

Table 4 shows which word classes are occurred in front of NE. This table indicates that outer context of NE is important clue for NER. About 50% of outer context is related to the noun class.

Table 4. Priori morphemes of NE

Sub-entity	percent
AE	3.7%
CE	9.8%
IE	4.2%
nb	0.8%
nc	11.0%
nn	1.5%
np	0.1%
sn	1.4%
SOS	15.6%
etc	52.0%

3. Named Entity Recognition Model

3.1 Supervised Learning Model

In this paper, we used a trigram in the learning phase. We extracted NE and prior morphemes and posterior morphemes in the NE tagged corpus. Following Fig. 1 describes a trigram model for variable length of NE[5]. L (Left) and R (Right) means prior and posterior morphemes respectively and n is the length of NE.

$$\cdots m_{-2}^{L} m_{-1}^{L} \quad m_1^{NE} m_2^{NE} \cdots m_{n-1}^{NE} m_n^{NE} \quad m_1^{R} m_2^{R} \cdots$$

Figure 1. Trigram model for variable length of NE

Given a sentence $W = w_1 w_2 w_3 \ldots w_m$, NER finds an optimal sequence $C = c_1 c_2 c_3 \ldots c_m$ that maximizes Pr(C|W). We implement a HMM[4] to

[1] In this paper, SE refers to IE, CE, and AE all together.

estimate Pr(C|W). Here is a HMM for Korean NER used our system.

N, State: a set of state S = $\{S_1, S_2, \ldots, S_N\}$. In the POS tagging system, a state corresponds to a POS and a state corresponds to a NE type in NER system. In normal, N is the number of NE type + 1(Not NE) and q_t means a state S in the time t. But, we modified the model by classifying "Not NE" with POS for discrimination. So, possible N is 37 (28 POS + 3 PLO * 3 SE).

- ***M*, Symbol**: observation symbol $V = \{v_1, v_2, \ldots, v_M\}$, In the NER, the number of observable symbol is the same as a size of the dictionary.

- ***A*, Transition Probabilities**: $A = \{a_{ij}\}$ is defined as follows:

$$a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i), 1 \le i, j \le N$$

- ***B*, Observation Probabilities**: $B = \{b_j(k)\}$ is defined as follows:

$$b_j(k) = P(q_t = v_k \mid q_t = S_i), 1 \le j \le N, 1 \le k \le M$$

- ***π*, Initial Distribution**: In the NER, π is the probabilities of some NE type i's occurrence in the start of the sentence. $\pi = \{\pi_i\}$ is defined as follows:

$$\pi_i = P(q_1 = S_i), 1 \le i \le N$$

Fig. 2 shows possible trigrams in the training corpus. We used typed trigrams which are encoded by **set(v_k)** function.

$$m_{-2}^{L}\, m_{-1}^{NE}\, m_{1}^{NE}$$
$$m_{-1}^{L}\, m_{1}^{NE}\, m_{2}^{NE}$$
$$\ldots$$
$$m_{n-1}^{NE}\, m_{n}^{NE}\, m_{1}^{R}$$

$$m_{n}^{NE}\, m_{1}^{NE}\, m_{2}^{R}$$

Figure 2. Extracted trigram

The function ***set(v_k)*** returns a sub-entity type for input morpheme v_k. In this paper, trigram is not a lexical level but a encoded type level such as "CE_location", "AE_person", or POS.

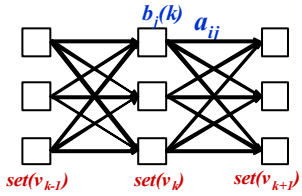

Figure 3. Transition and observation probabilities

We calculate the transition and observation probabilities as follows:

Transition Probabilities

$$a_{ij} = p(set(v_k) \mid set(v_{k-1}), set(v_{k+1}))$$
$$= \frac{p(set(v_{k-1}), set(v_{k+1}) \mid set(v_k))\, p(set(v_k))}{p(set(v_{k-1}), set(v_{k+1}))}$$
$$\cong \frac{p(set(v_{k-1}) \mid set(v_k))\, p(set(v_{k+1}) \mid set(v_k))\, p(set(v_k))}{p(set(v_{k-1}), set(v_{k+1}))},$$

$$p(set(v_{k+1}) \mid set(v_k)) = \frac{\sum_{x=1}^{N} p(set(v_k)_j, set(v_{k+1})_x)}{\sum_{x=1}^{N} \sum_{x=1}^{N} p(set(v_k)_x, set(v_{k+1})_x)}$$

Observation Probabilities:

$$b_j(k) = p(q_t = v_k \mid q_t = S_j), 1 \le k \le M$$
$$b_j(k) = p(v_k \mid set(v_k)) = \frac{p(set(v_k) \mid v_k)\, p(v_k)}{p(set(v_k))}$$

3.2 Features for NER

We only used two kinds of features. First is the combination of sub-entity and NE types for classification. Second is the combination of status, sub-entity, and NE types for detection.

- **classification feature examples**

  - *seoul*: IE_location
  - *kong-sa*: CE_organization, IE_organization
  - *dan-won*: AE_person

- **detection feature examples**

  - *cen-ca*: CE_organization_end, CE_organization_continue
  - *Kim-dai-cung*: IE_person_unique
  - *sang-sa*: CE_organization_end, AE_person_not

As an example of NE detection, following encoded feature sequence is possible. By our NER model, we can detect that an organization name covers B, C and D.

A: AE_organization_not
B: IE_organization_start
C: nc_continue
D: CE_organization_end
E: jxc_not

With these features, we propose a supervised learning method for NER that uses the relationship among IE, CE, AE and other nouns. This relationship is encoded with the distribution probabilities among them.

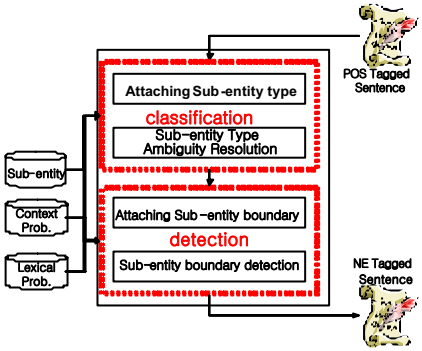

Figure 4. Structure of the NE Recognition System

Our method has two phases. First phase is a classifying step. All words are tagged with possible IE, CE and AE or POS (if a word is not included in an IE, CE or AE, that will be tagged with a POS). A word can have multiple entity types. For examples, "*ceng-sen*" will be a person name or a location name in Korean. "*sang-sa*" will be an AE-person (a master sergeant) or a CE-organization (business affairs). Second phase is a detecting step that finds the boundaries of the NE and identifying the NE type (PLO).

4. Experimental results

We gathered about 68,000 person, 25,000 location, and 10,000 organization name for constructing an IE dictionary. Also, we have a CE dictionary with 92 location and 121 organization entries and AE with 114 person, 39 location and 33 organization entries.

Table 5. NE Dictionaries

	IE	CE	AE
# of Entries	102,841	213	186

We evaluated our system with two different unseen data sets. Each data set is consisted of 10 articles. Following table describes the results of our system.

Table 6. Experimental results

	Recall	Precision	F-measure[2]
Performance articles	81.5%	86.3%	83.8%
Economic articles	84.2%	90.1%	87.1

The major types of errors are shown below:

- common words

  - Sometimes, common words rarely used as

[2] $F-measure = \frac{2*P*R}{P+R}, P: \Pr ecision, R: \Re call$

a NE. For example, "*ku-lim* (a picture)" may be a person name like "Grimm Brothers", or "*gyeng-sang* (an ordinary in the economics term **ordinary loss**)" be used as a location.

- CE combination
  - "*kuk* (a bureau)" is a CE word. But, all the words ended with a "*kuk*" are regarded as an organization name in our system. But, "*kong-up-kuk* (an industrial country)" is not an ORG name.

These errors indicate that the system needs mutually exclusive word-level information. So, our future works included such works.

5. Conclusions

Previous works for Korean NER are mainly with rule-based methods. Such systems need a human-labored works for domain portability. For overcoming these problems, we proposed a machine learning based approach. We used the HMM with the compound word construction principles for classifying and detecting a NE. The transition probability is calculated by the encoded trigram and observation probability is calculated by the lexical frequency of each entities NE from labeled training data. Experimental result shows that this method is a reasonable approach for Korean NER.

Our future works included the integration of classification and detection phase by learning trigram with a maximal information and extension of SE dictionary.

6. References

[1] D. M. Bikel, S. Miller, R. Schwartz, R. Weischedel, "Nymble: A High-Performance Learning Named-finder", In Proceedings of the Fifth Conference on Applied Natural Language Processing, pp.194-201, 1997.

[2] M. Collins and Y. Singer, "Unsupervised Models for Named Entity Classification", EMNLP/VLC-99, pp. 189-196, 1999.

[3] K. H. Lee, J. H. Lee, M. S. Choi, G. Ch. Kim, "Study on Named Entity Recognition in Korean Text", In Proceedings of the 12th Hangul and Korean Information Processing, pp. 292-299, 2000.

[4] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, Vol. 77, No. 2, pp. 257-286, 1989.

[5] M. Sassano and T. Utsuro, "Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition", Proceedings of the 18th International Conference on Computational Linguistics, pp.705-711, 2000.

[6] S. Sekine, R. Grishman and H. Shinnou, "A Decision Tree Method for Finding And Classifying Names in Japanese Texts", Proceedings of the Sixth Workshop on Very Large Corpora, 1998.

[7] C. N. Seon, Y. Ko, J. S. Kim, and J. Seo, "Named Entity Recognition using Machine Learning Methods and Pattern-Selection Rules", NLPRS 2001, pp. 229-236, 2001.

[8] K. Uchimoto, Q. Ma, M. Murata, H. Ozakum, and H. Isahara, "Named Entity Extraction Based on A ME Model and Transformation Rules", In Processing of the ACL 2000.

[9] J. Fukumoto, M. Shimohata, F. Masui, and M. Saski, "Description of the Oki System as Used for MET-2", In Proceedings of 7th Message Understanding Conference, 1998.