World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:8, 2007

Determining the Gender of Korean Names for Pronoun Generation

Seong-Bae Park and Hee-Geun Yoon

***Abstract*—**It is an important task in Korean-English machine translation to classify the gender of names correctly. When a sentence is composed of two or more clauses and only one subject is given as a proper noun, it is important to find the gender of the proper noun for correct translation of the sentence. This is because a singular pronoun has a gender in English while it does not in Korean. Thus, in Korean-English machine translation, the gender of a proper noun should be determined. More generally, this task can be expanded into the classification of the general Korean names. This paper proposes a statistical method for this problem. By considering a name as just a sequence of syllables, it is possible to get a statistics for each name from a collection of names. An evaluation of the proposed method yields the improvement in accuracy over the simple looking-up of the collection. While the accuracy of the looking-up method is 64.11%, that of the proposed method is 81.49%. This implies that the proposed method is more plausible for the gender classification of the Korean names.

***Keywords*—**machine translation, natural language processing, gender of proper nouns, statistical method.

I. Introduction

IN Korean an omission of subject or object which is considered as a complement freely occurs, where this is in general impossible in European languages such as English, French, German, and so on. Especially, the omission of a subject makes a severe problem in the machine translation systems of which target language requires a subject as a complement. For instance, in Korean-English machine translation system, a sentence without a subject in the source language, Korean, is valid, but the corresponding sentence translated into English without the subject is ungrammatical.

When a sentence is composed of two or more clauses and only one subject is given as a proper noun, there are two possibilities whether or not the clause with a subject shares the subject with those without a subject. When a subject is shared by multiple clauses, it is important to determine the gender of the subject in English-Korean machine translation. This is because a different pronoun should be generated for the clauses without a subject according to the gender of the shared subject. The goal of this paper is, therefore, to determine the gender of the subject. More generally, this task can be expanded into a classification of normal Korean names. However, it is a totally independent problem to know whether a subject is shared by several clauses or not. Thus, in this paper, we just focus on the gender classification of Korean names.

Figure 1 shows an error of the Google translator related with our topic. For the input sentence “영희는 극장에 가서 영화를 봤다.”, the Google translator outputs “The zero [huy] went to the teater and it saw a movie.”, where the correct sentence is “Young-Hee went to movie.” Even though the translation is made in the surface level, it should be “Young-Hee went to the theater and see a movie.” The first error is made in failing in the named entity recognition. ‘영희 (Young-Hee)’ is actually a name of an female, but it is not recognized as a name in the Google translator. As a result, ‘영희’ is translated into ‘zero [huy]’. The worse is that ‘영희’ is translated into ‘it’ in the second clause. This is because the Google translator did not know whether ‘영희’ is human or not and it is female or not.

In order to tackle this problem, two tasks have to be solved. The first task is named-entity recognition. Each proper noun representing human names should be correctly recognized as a name. This is a relative well-known problem in natural language processing community, and has been widely and deeply studied for last a few years. The recent work in this task reports very high performance [13].

The second task is to determine the gender of the recognized names. Only when the gender of a name is correctly determined, the proper pronoun can be generated in the machine translation systems. To solve this task, a probabilistic method is adopted in this paper, where the probabilistic information on names is gathered from open-domain web pages. Since the named entity recognition is independent from the determination of the gender for a given name and has been studied by various researchers, this paper focuses only on this task.

The rest of this paper is organized as follows. Section 2 surveys the previous work on named-entity recognition and pronoun generation in Korean. Section 3 proposes how the gender of Korean names is classified using a statistical method, and Section 4 gives an overall structure how the proposed method is used in Korean-English machine translation. Section 5 presents the experimental results. Finally, Section 6 draws conclusions.

II. Related Work

The named-entity recognition has been implemented by two kinds of methods. The first one is rule-based methods which use regular-expression-like patterns and dictionaries [11]. When the dictionaries contain the great number of entities and the general patterns are extracted from a large-scale corpus, the performance of rule-based methods will be good. However, the extraction of rules is in general very difficult, and the cost of constructing such large dictionaries is extremely high.

The other is to use the statistical information for this task. This methods collect the statistical knowledge from entity-

Seong-Bae Park and Hee-Geun Yoon are with Department of Computer Engineering, Kyungpook National University, Daegu 702-01, Korea.

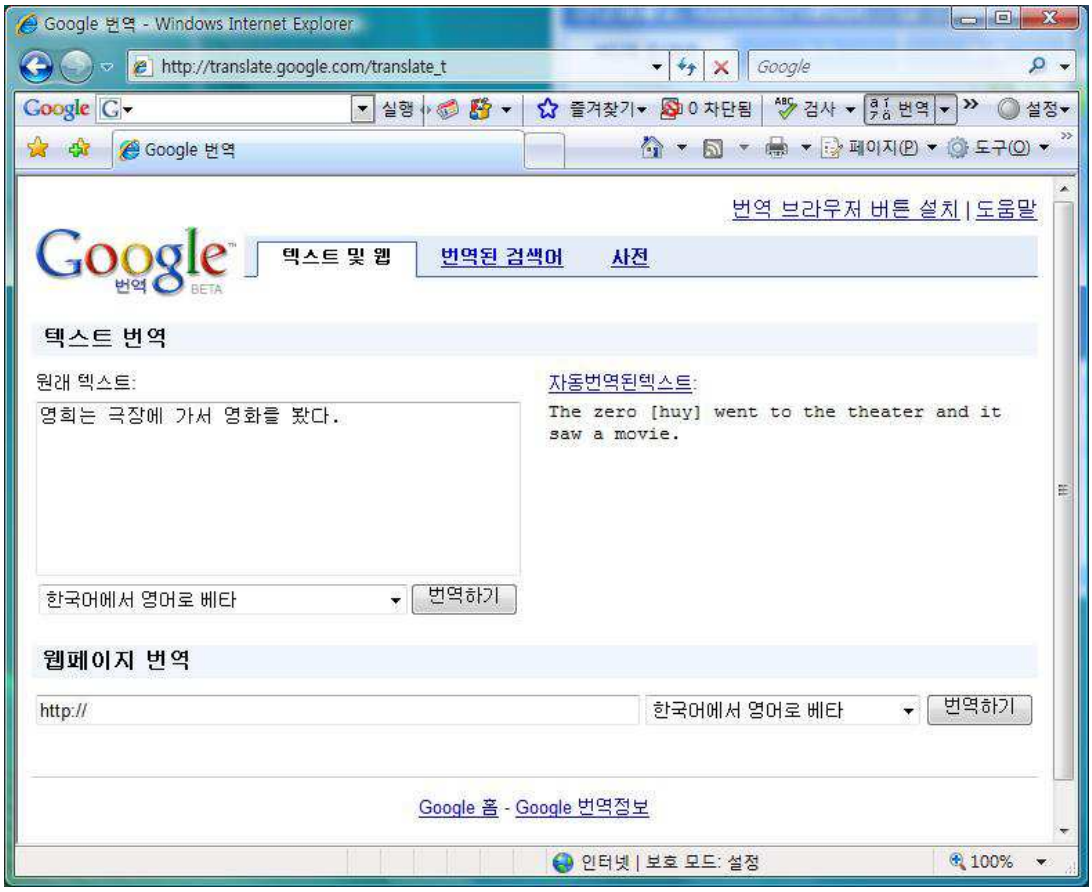

Fig. 1. An example mistake which the Google translator makes.

tagged corpora, and then determine the entity category using the knowledge. Since there is no large-scale entity dictionary for Korean, most named entity recognizers for Korean uses the statistical methods [7]. What determines the performance of these methods is the size of tagged corpora [1]. This is because the large scale corpora reduce the possibility of data sparseness. As a solution of this phenomenon, various types of semi-supervised methods have been adopted. For instance, Chung et. al [1] and Kwak [7] use the co-training algorithm, while Kim et. al propose an acyclic weighted digraph [5].

The generation of an appropriate pronoun is a classical problem in Korean-English machine translation. This problem is well-known as 'zero-pronoun problem'. The zero-pronoun is a common problem in both Korean and Japanese and it is considered as a very important problem in corpus annotation [9] and Korean language processing [3] Roh and Lee adopt a linguistic theory called 'centering theory' to solve this problem [10]. While machine learning methods are common in this problem [12], they are not widely used for processing zero pronoun in Korean. As a first step to solve zero pronoun problem in Korean, we first focus on the gender determination of Korean names.

III. Name as a Sequence of Syllables

From the point of view for machine learning, the task of classifying the gender for a given name is a kind of classification task. That is, for a given name N, the task is to determine its gender $y \in Y = \{male, female\}$. When a training set $D = \{(N_1, y_1), (N_2, y_2), \ldots, (N_n, y_n)\}$ composed of a pair of a name and its gender, the task is to find a function $f : N \rightarrow Y$.

Such a function f can have various forms, the conditional probability is one of simplest and strongest candidates for the function f. That is, if the conditional probability is used for f, the task which we focus on is to determine the gender of a newly incoming name N' as follows.

$$y* = \underset{y \in \{male, female\}}{\arg\max} P(y|N', \theta), \quad (1)$$

where θ is a parameter for the distribution P and it is determined from the training data D.

The Korean names are in general composed of $2 \sim 5$ consecutive syllables. That is, $N = x_1, x_2, \ldots, x_m$, where the maximal length of a name is m and each syllable within a name is x. Therefore, the probability of the the name N is

$$\begin{aligned} P(N) &= P(x_1, x_2, \ldots, x_m) \\ &= P(x_1)P(x_2|x_1)\ldots P(x_m|x_1, \ldots, x_{m-1}). \end{aligned} \quad (2)$$

If we approximate $P(x_i|x_1, \ldots, x_{i-1})$ by its count, it is rewritten as

$$P(x_i|x_1, \ldots, x_{i-1}) = \frac{C(x_1, \ldots, x_i)}{C(x_1, \ldots, x_{i-1})} \quad (3)$$

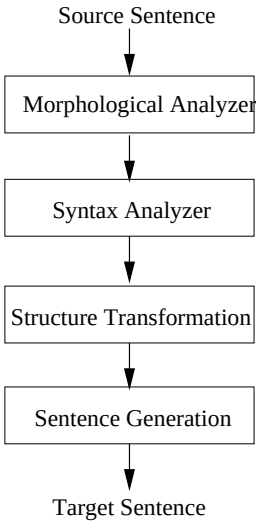

Fig. 2. The general steps in machine translation systems.

where C is a counting function.

Since we consider a name as a sequence of syllables, it is considered to be a n-gram model. That is, a long name implies high dimensionality in this model. The probability in Equation (3) is easy to be 0 in high dimensionality. It leads to making $P(N)$ in Equation in (2) zero. As a result, it gets impossible to distinguish the gender of a given name using probability. In order to solve this problem, this paper adopts the Back-off model proposed by Katz [4].

IV. Overall System

When a module for determining the gender of names, it can be used as a subsystem for generating pronouns in Korean-English machine translation. Generally, as shown in Figure 2, the machine translation system consists of four steps: morphological analysis, syntax analysis, structure transformation, and target language generation [6]. Thus, the proposed method is used in the target language generation step.

For a given sentence, it first distinguishes the named entities. As a named-entity recognizer, we use the one developed in ETRI [8]. For each named entity found by ETRI NER, if it is a person, its gender is determined by the proposed method. According to the found gender, an appropriate pronoun is generated. Figure 3 depicts this process.

The "Determine Pronoun" step in Figure 3 is shown in more detail in Figure 4. Since this step is reached only when the named entity is people, the target pronoun p can be one of *he*, *she*, and *then*. The first step is to look up the result of morphological analyzer in order to see if it is plural or not. If this entity is plural, p should be *then*. Otherwise, p is *he* when g is male. It is *she* when g is female.

V. Experiments

A. Data Sets

For the evaluation of the method proposed in this paper, we collected names on personage databases which are available publicly on Web. The number of names collected in this way

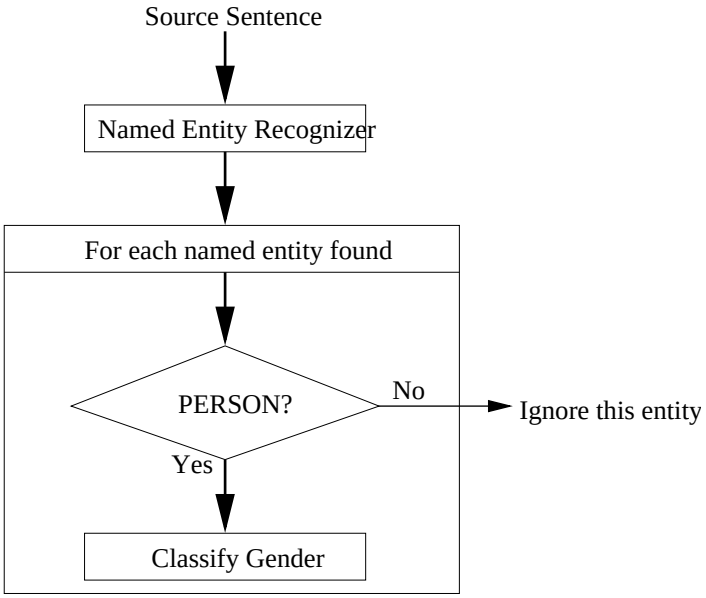

Fig. 3. The process of pronoun generation for proper nouns in Korean-English machine translation.

Procedure Determine Pronoun
Input : a name N and its gender g
Output : a pronoun p

[Step 1] num = Look up the number of N.
[Step 2] **if** num = *plural* **then** p = 'they'
[Step 3] **elseif** g = *male* **then** p = 'he'
[Step 4] **else** p = 'she'.

Fig. 4. The procedure for generating an appropriate pronoun for a human name with its gender.

is . Among them, are man's name. That is, 97.02% of the generally well-known people are men. This reflects Korean social characteristics that the social activity of women is weak.

However, it also causes a problem in classifying the gander of names by probability, since the imbalanced data distribution hurts the performance of most machine learning algorithms [2] In order to solve this problem, we manually collected additional 23,488 female names from Web. Finally, we could obtain the balanced data set, and Table I shows the simple statistics on this data set. The number of names in this table is obtained after removing the redundant ones. For instance, the number of female name is 23,923, but it shrinks to 12,145 after removing the redundant ones.

B. Compared Methods

In order to show the performance of the proposed method, we compare it with three other methods explained below. The easiest way to determine the gender of a name is to reply the majority gender in the corpus. According to Table I, when the system replies "Male" for any name, it will achieve 52.24% of accuracy. This is our baseline performance for this task.

The other simple method is to look up the name database. However, this method has two kinds of problems. First, it

TABLE I
THE SIMPLE STATISTICS ON THE NAME DATA SET.

Gender	No. of Instances	Percentage (%)
Male	13,284	52.24
Female	12,145	47.76

TABLE II
THE EXPERIMENTAL RESULTS OF THE PROPOSED METHOD FOR DETERMINING GENDER OF NAMES.

Method	Accuracy (%)
Baseline	52.24
Lookup	64.11 $\pm$ 0.46
Trigram	75.27 $\pm$ 1.89
Proposed Method	81.49 $\pm$ 1.56

will fail to find the gender of a name when the name is not listed in the database. Since there is no law to restrict naming, it is always possible to miss a name in the database. The second problem is that some names can be for both man and woman. When a family name is given, the probability to find the gender for the names gets higher. However, without the contextual information, their gender can not be determined in high accuracy.

The third and more complex one is to use statistical information like the proposed method. Rather than using whole sequence of names, this method is implemented by the trigram model. That is, each $P(x_i|x_1,\ldots,x_{i-1})$ in Equation (2) is assumed to be

$$P(x_i|x_1,\ldots,x_{i-1}) \approx P(x_i|x_{i-2},x_{i-1}).$$

Thus, Equation (2) is simply computed by

$$\begin{aligned} P(N) &= P(x_1,x_2,\ldots,x_m) \\ &= P(x_1)P(x_2|x_1)\ldots P(x_m|x_1,\ldots,x_{m-1}) \\ &= \prod_{i=1}^{m} P(x_i|x_{i-2},x_{i-1}). \end{aligned}$$

The benefit of this method is that it can achieve more stable probability from a limited-sized corpus. However, its expressive power is weaker than that of the proposed method.

C. Experimental Results

Table II shows the experimental results of the proposed method. 'Baseline', 'Lookup', and 'Trigram' in this table are the ones explained in the previous section. All experiments for 'Lookup', 'Trigram', and 'Proposed Method' are done in 5-fold cross validation. 'Baseline' reports 52.24% of accuracy, 'Lookup' 64.11%, 'Trigram' 75.27, but our 'Proposed Method' achieves 81.49%. That is, the proposed method improves up to 29.25% of accuracy. It even outperforms 'Trigram' by 6.22%.

Figure 5 shows the effect of the number of names in this task. The X-axis in this figure is the number of names used, and the Y-axis is the accuracy. As shown in this figure, the accuracy increases monotonically as the number of names increases. Even though the increase slope gets slow after around 7,000 names, the performance increase is expected with more names. It implies that the proposed method is plausible with the reasonable size of name collection.

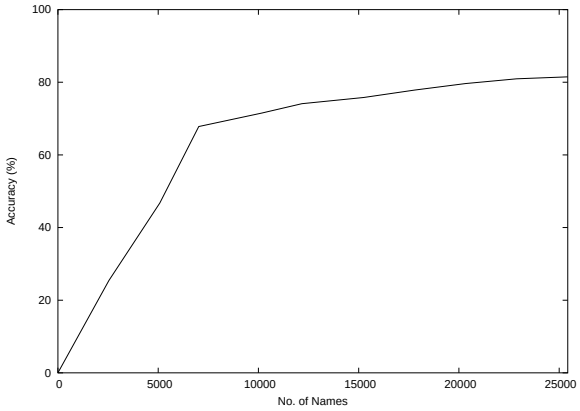

Fig. 5. The effect of the number of names in the gender classification of names.

VI. CONCLUSION

In this paper, we have proposed a method for determining the gender of Korean names for Korean-English machine translation. The proposed method is base on the statistics computed with a collection of Korean names. According to the experiments, the proposed method gives 81.49% of accuracy, while the database looking-up achieves just 64.11%. That is, the proposed method improves 17.38% of accuracy over the simple looking-up. It implies that the proposed method is plausible for gender classification of Korean names in Korean-English machine translation systems.

ACKNOWLEDGMENT

This work was supported by the Korean Ministry of Education under the BK21-IT Program.

REFERENCES

[1] E.-S. Chung, Y.-G. Hwang, and M.-G. Jang, "Korean Named Entity Recognition Using HMM and Co-Training Model," In *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages*, pp. 161–167, 2003.
[2] C. Drummond and R. Holte, "C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling," In *Proceedings of Workshop on Learning from Imabalanced Datasets II, ICML*, 2003.
[3] N.-R. Han, *Korean Zero Pronouns: Analysis and Resolution*, Ph.D Thesis, University of Pennsylvania, 2006.
[4] S. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 35, No. 3, pp. 400–401, 1987.
[5] K.-N. Kim, Y.-H. Yoon, H.-S. Kim, and J.-Y. Seo, "Named Entity Recognition Using Acyclic Weighted Digraphs: A Semi-Supervised Statistical Method," *Lecture Notes in Computer Science*, Vol. 4426, pp. 571–578, 2007.
[6] Y.-T. Kim, *Introduction to Natural Language Processing*, 2nd Edition, Saeng-Neung Publisher, 2001. (*In Korean*)
[7] B.-K. Kwak and J.-W. Cha, "Named Entity Tagging for Korean Using DL-CoTrain Algorithm," *Lecture Notes in Computer Science*, Vol. 3689, pp. 589–594, 2005.
[8] C.-K. Lee, Y.-G. Hwang, H.-J. Oh, S.-J. Lim, J. Heo, C.-H. Lee, H.-J. Kim, J.-H. Wang, and M.-G. Jang, "Fine-Grained Named Entity Recognition Using Conditional Random Fields for Question Answering," *Lecture Notes in Computer Science*, Vol. 4182, pp. 581–587, 2006.
[9] S.-H. Lee, D. Byron, and S.-B. Jang, "Why Is Zero Marking Important in Korean?" In *Proceedings of the 2nd International Conference on Natural Language Processing*, pp. 588–599, 2005.

[10] J.-E. Roh and J.-H. Lee, "Generation of Zero Pronouns Based on the Centering Theory and Pairwise Salience of Entities," *IEICE Transactions on Information and Systems*, Vol. E880D(2), pp. 837–846, 2006.
[11] C.-N. Seon, Y-.J. Ko, J. Kim, and J.-Y. Seo, "Named Entity Recognition Using Machine Learning Methods and Pattern-Recognition Rules," In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, 2001.
[12] S. Zhao and H. Ng, "Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach," In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 541–550, 2007.
[13] G. Zhou and J. Su, "Named Entity Recognition Using an HMM-Based Chunk Tagger," In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 473–480, 2002.