A Multi-phase Semi-supersense Tagging of Korean Unknown Nouns

Young-Bum Kim[1], Jung-Kuk Lee[2,3], and Yu-Seop Kim[2,3,*]

[1] Dept of Computer Science, University of Wisconsin-Madison,
1210 W. Dayton St. Madison, WI 53706-1685
stylebbum@gmail.com
[2] Dept. of Ubiquitous Computing, Hallym University, 1 Hallymdaehak-gil,
Chuncheon, Gangwon-do, 200-702 Korea
[3] Bio-IT Research Center, Hallym University, 1 Hallymdaehak-gil,
Chuncheon, Gangwon-do, 200-702 Korea
yskim01@hallym.ac.kr

Abstract. Supersense tagging is a problem of finding a corresponding semantic super tag (eg. Phenomenon, Act) based on syntactic information and annotated corpora. However, we employ semantic information rather than syntactic one and annotated corpora, because Korean language has relatively flexible syntactic structure and is lack of annotated corpora. To construct the automatic sense tagging system for Korean language, we use semi-supersenses of first and second level in Sejong's Noun Semantic Class System. We employ a hybrid approach consisting of three phases: one morphological matching phase and two semantic matching phases. The morphological phase is based on suffix pattern matching which assigns compound word to the class including the suffix word. One of the two semantic matching phases is based on concept similarity on WordNet, and the other is based on the term similarity in term matrix reduced by singular value decomposition (SVD). Above semantic phases are using weighted k-Nearest Neighbor classifier commonly but are also using different similarity metrics. In experiments, 79,103 unknown words are extracted from 225,779 noun words from syntactic tagged corpus, and 98% of the unknown words are addressed by our hybrid method.

1 Introduction

With the advancement and growth of Semantic Web, it is required to construct lexical-semantic resources by converting from unstructured information expressed in raw texts to structured view [1]. Approaches dealing with these tasks are Named Entity Recognition (NER), identification and classification of words in a corpus with a proper named entity type (People, Place and Organization) [2], and a Super Sense Tagging (SST) to extend the NER system [3].

* Corresponding author.

G. Lee et al. (Eds.): ICHIT 2012, CCIS 310, pp. 760–766, 2012.

Of two systems, the SST systems are recognized as better systems because of higher recognition performance followed by broad coverage [4]. These SST researches up to now can be divided into two approaches: supervised and unsupervised ones. The one of the supervised approaches is a multi-class perceptron tagger [3]. This tagger uses synonym set glosses in the Wordnet [5] as annotated training data. The unsupervised approach brings the tagger created by [6]. This system uses vector space similarity based on the SEXTANT system [7].

These tagging systems have been developed not only for English, but also for other languages such as Italian [8] and Chinese [9].

Contrast to English tagger, the SST systems for other languages use different metrics and training sets because each language has different syntactic properties and morphological features [10].

Similarly, in this paper, we present a new method for the supersense tagging for Korean. We should find a new method because of the following three reasons. First, Korean has limitation of using syntactic information because of relatively free word order, and thus its relatively flexible syntax. Second, we do not have thesaurus such as Wordnet, which contains well-organized glosses. Finally, Korean has its own Sejong semantic class system of noun.

What we propose is a hybrid approach consisting of three phases: one morphological matching phase and two semantic matching phases. The morphological phase is based on suffix pattern matching which assigns a compound word to the class including the suffix word itself. And the semantic matching phase has two sub-phases. One semantic matching phase is based on concept similarity on WordNet, and the other is based on the term similarity in term matrix reduced by singular value decomposition (SVD). Above semantic phases are all using k-Nearest Neighbor classifier commonly but are also using different distance metrics. The concept similarity and the term similarity are used for the metric. We calculate the weighted sum of the k nearest words' distances and decide the class, which is the semi-supersense, with the highest sum as the target class.

The rest of this paper is structured as follows: Section 2 explains what the semi-supersense tagging is. Section 3 describes multi-phase semi-supersense tagging. Section 4 shows experiments and evaluation of our approach and concluding remarks are discussed in section 5.

2 Semi-supersense Tagging

Semi-Supersense Tagging is to extend the English supersense system based on Wordnet. In order to construct semi-supersenses, we extract semantic classes of first and second level from Sejong semantic class system of noun.

Generally, English SST systems [3][6] employ 26 supersenses, "lexicographer class" labels used in WordNet. The supersense lables that WordNet lexico-graphers use to organize nouns are listed in Table 1 [11].

Table 1. Supersenses in WordNet

Act, animal, artifact, attribute, body, cognition, communication, event, feeling, food, group, location, motive, object, quantity, phenomenon, plant, possession, process, person, relation, shape, state, substance, time, tops

However, because the number of concept is too small, a single supersense includes too many concepts, which makes many terms losing its polysemous characteristics, and it is difficult to distinguish terms from concepts [12]. Therefore, we replace the 26 supersenses in WordNet with 74 semi-supersenses, semantic classes extracted from the Sejong semantic class of noun described in Table 2. The semi-supersense is going to be the target sense when we tag the given noun terms.

Table 2. This table shows supersenses and semi-supersenses in each supersense in Sejong semantic noun class

Supersense	Semi-supersense	# of Semi-supersense
Physical Object	Physical natural object, Physical artifact, and etc.	4
Group	Human group, Non-Human group	2
Place	Ground place, Water place, and etc.	14
Abstract Object	Money, Time, Method, Skill, Role, and etc.	44
State	Static state, Act, Incident, Phenomenon, and etc.	5

3 Multi-phase Semi-supersense Tagging

This section describes our multi-phases approach for semi-supersense tagging of Korean unknown words, including one morphological matching phase and two semantic matching phases.

Fig. 1 shows the whole process of the multi-phases for the sense tagging of unknown nouns. 'Out of vocabulary (OOV)' means the nouns not included in semi-supersenses of Sejong semantic class. If the OOV passes through each matching stage and is annotated by its sense, it is registered in the semi-supersense table.

3.1 Morphological Matching Phase

The morphological matching phase is the first stage to deal with OOVs that do not appear in semi-supersenses. For morphological matching, we find specific patterns of nouns included in each semi-supersenses.

In case of Korean compound words, a former noun generally qualifies a latter noun. Namely, a compound noun is a very specific form the latter noun. For example as 'Min-Ju-Ju-Wi (in English, 'Democracy') is a specific form of 'Ju-Wi' (in English, 'belief'). Fig. 2 shows how the suffix 'Ju-Wi' is appeared in compound nouns classified in Gai-Num (in English, 'concept') semi-supersense.'

Of unknown words, a word which has a pattern of a certain semi-supersense is assigned into equivalent semi-supersense. For example, 'Dang-Pa-Ju-Wi' (in English, 'exclusivism') has a 'Ju-Wi' pattern of the 'Gai-Num' semi-supersense, and then it is assigned into 'Gai-Num' semi-supersense.

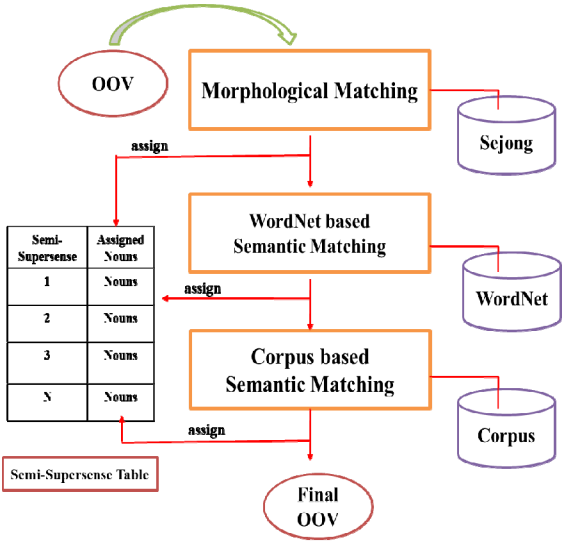

Fig. 1. The whole process of tagging phases

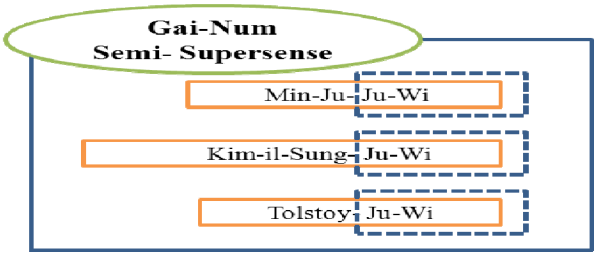

Fig. 2. The example of a pattern in Gai-Num semi-supersense

3.2 Semantic Matching Phase

In this section, we describe two semantic matching phases: WordNet based one and corpus based one. We predict the semi-supersense of an unknown word by using a k-Nearest Neighbor (KNN) classifier.

Firstly, in order to calculate similarity between concepts in WordNet, we revise [13]'s formula to

$$Sim(c_1, c_2) = -\log \frac{len(c_1, c_2)}{2D} + k,$$

where D means depth of LSO(Lowest Super-Ordinate), lowest depth synset including both concepts together, $len(c_1, c_2)$ is the number of nodes between concepts because of prevention of log zero operation exception and k is a constant value which enables all value of similarity to make a positive number, and then obtain correct weight summation. In this paper, k is 1 because lowest value of similarity between concepts in WordNet is -0.954.

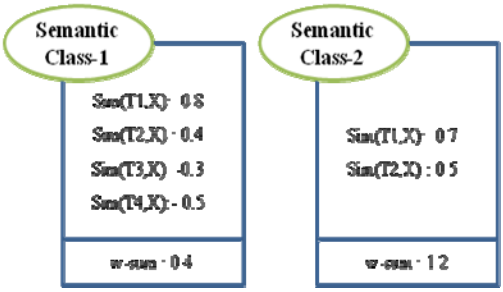

Fig. 3. An exceptional case when k=0

Fig. 3 is an exceptional case when k is zero. Specifically, although semantic class-1(SC1) has more similar words than semantic class-2(SC2), a certain word X is assigned to the latter class due to the fault of calculation of weight sum in SC1.

Second, in order to calculate similarity between terms in corpora, we construct term similarity matrix reduced by singular value decomposition (SVD). We extract n indexed terms from N documents. A document d_j has n dimensions and is represented by.

$$d_j = \langle w_{1,j}, w_{2,j}, \ldots, w_{i,j}, \ldots, w_{n,j} \rangle$$

where $w_{i,j}$ denotes a weight of ith element of the document d_j. To take a consideration of both global and local information, the following is the computation of weight:

$$W_{i,j} = tf_{i,j} \cdot idf_i = tf_{i,j} \cdot \log\left(\frac{N}{df_i}\right) \tag{1}$$

In formula (1), $tf_{i,j}$ stands for occurrence frequency of the ith element in the given document d_j, df_i, called as a document frequency, denote the number of documents including the ith element, and idf_i is an inverse document frequency.

An initial corpus matrix M_c is constructed by considering every single d_j vector as a column vector of M_c. Then, the row vector of M_c represents each indexed term vector,

$$t_j = \langle w_{1,j}, w_{2,j}, \ldots, w_{i,j}, \ldots, w_{n,j} \rangle$$

With the term-document matrix M_c that represents term knowledge in a semantic network as a sparse matrix, we build the latent semantic kernel, using singular value decomposition(SVD), in order to reduce its dimensionality, obtaining the most important correlation in that knowledge. The initial matrix M_c is then transformed into P which is used in the following formula [14].

$$Sim(t_1, t_2) = \cos\left(P^T d_1, P^T d_2\right) = \frac{d_1^T P P d_2}{\left|P^T d_1\right|\left|P^T d_2\right|}$$

4 Experiment and Evaluation

4.1 Data Set

As experiment data sets, we use 225,779 common nouns, which are extracted from a part-of speech tagged corpus created by Electronics and Telecommunication Research Institute [15], and the corpus has 101,602 sentences.

We firstly eliminate special symbols except for '_' identifier which separate tokens in compound nouns and not-Korean language such as pure Chinese words, English words and numbers, and then use 197,225 nouns, not considering repetitive words. Second, we separate compound nouns from common ones. The number of compound and single nouns is 137,169 and 60,056 respectively.

Both single and compound nouns basically are assigned into semi-supersenses through exact matching words stage, but when processing compound nouns, we compare the ending segments word in compound words with member words in the semi-supersenses. Finally, the number of unknown words is 46,289 in case of common words, and 32,184 in case of compound words. For corpus based semantic matching, we use news article 38,000 documents and the number of type is 280,038.

4.2 Performance Evaluation

In order to measure the accuracy or our approach, we randomly select different 100 nouns which are processed by each stage of matching, and a human assessor checks whether assigned a semi-supersense is correct. Table 3 shows the accuracy of each matching stage.

Table 3. Accuracy of each matching stage

Matching stage	# of wrong matching	Accuracy rate
Pattern	3	97%
WordNet	7	93%
Corpus	20	80%

Table 4 shows the ratio of solved terms in each stage. The left one is for the single words and the right one is for the compound words.

Table 4. The coverage of each stage in tagging process

Matching stage	# of solved OOV		Ratio of solved OOV	
Pattern	26756	11705	57%	36%
WordNet	4870	9527	11%	30%
Corpus	14663	10360	32%	34%

5 Conclusion

We describe multi-phase semi-supersense tagging, consisting of one morphological matching phase and two semantic matching ones. Although the characteristics of Korean language causes difficulties regarding tagging unknown words, our approach shows good tagging performance in terms of accuracy and coverage. Furthermore, this technique is helpful for constructing semantic resources. However, some unknown nouns such as proper nouns was not exactly assigned into semi-supersenses. Therefore, in the future work, we will construct an ensemble model with various domain-specific corpora.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation(NRF) funded by the Ministry of Education, Science and Technology(2010-0010612).

References

1. Picca, D., Popescu, A.: Using Wikipedia and supersense tagging forsemi-automatic complex taxonomy construction. In: RANLP 2007, CALP Workshop (2007)
2. Collins, M., Singer, Y.: Unsupervised Models for Named Entity Classification. In: The Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. (1999)
3. Ciaramita, M., Johnson, M.: Supersense Tagging of Unknown Nouns in WordNet. In: The 2003 Conference of Empirical Methods in Natural Language Processing, pp. 168–175 (2003)
4. Marrero, M., Sanchez-Cuadrado, S., Lara, J., Andreadakis, G.: Evaluation of Named Entity Extraction Systems. Advances in Computational Linguistics, Research in Computing Science 41, 47–58 (2009)
5. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
6. Curran, J.: Supersense Tagging of Unknown Nouns Using Semantic Similarity. In: ACL 2005, pp. 26–33 (2005)
7. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers, Boston (1994)
8. Picca, D., Gliozzo, A.M., Ciaramita, M.: Supersense Tagger for Italian. In: The Sixth International Conference on Language Resources and Evaluation (2008)
9. Lu, X.: Hybrid Models for Semantic Classificaiton of Chinese Unknown Words. In: NAACL HLT 2007, pp. 188–195 (2007)
10. Chen, K.-J., Chen, C.-J.: Automatic Semantic Classification for Chinese Unknown Compound Nouns. In: COLING 2000, pp. 173–179 (2000)
11. Ciaramita, M., Altun, Y.: Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger, Source. In: The Conference on Empirical Methods in Natural Language Processing (2006)
12. Sowa, J.: Knowledge Representation: Logical Philosophical and Computational Foundations. Brooks and Cole (1999)
13. Leaock, C., Chodrow, M.: Combining Local Context and WordNet Similarity for Word Sense Identification. In: Fellbaum, pp. 265–283 (1998)
14. Cristianini, N., Shawe-tayler, J., Lodhi, H.: Latent Semantic Kernel. Journal of Intelligent Information Systems 18(2-3), 127–152 (2002)
15. ETRI: POS Tag Guidelines. Technical Report, ETRI, Taejun, Korea (1999)