Automatic WordNet mapping using word sense disambiguation*

Changki Lee
Geunbae Lee[†]
Natural Language Processing Lab
Dept. of Computer Science and Engineering
Pohang University of Science & Technology
San 31, Hyoja-Dong, Pohang, 790-784, Korea
{leeck,gblee}@postech.ac.kr

Seo JungYun
Natural Language Processing Lab
Dept. of Computer Science
Sogang University
Sinsu-dong 1, Mapo-gu, Seoul, Korea
seojy@ccs.sogang.ac.kr

Abstract

This paper presents the automatic construction of a Korean WordNet from pre-existing lexical resources. A set of automatic WSD techniques is described for linking Korean words collected from a bilingual MRD to English WordNet synsets. We will show how individual linking provided by each WSD method is then combined to produce a Korean WordNet for nouns.

1 Introduction

There is no doubt on the increasing importance of using wide coverage ontologies for NLP tasks especially for information retrieval and cross-language information retrieval. While these ontologies exist in English, there are very few available wide range ontologies for other languages. Manual construction of the ontology by experts is the most reliable technique but is costly and highly time-consuming. This is the reason for many researchers having focused on massive acquisition of lexical knowledge and semantic information from pre-existing lexical resources as automatically as possible.

This paper presents a novel approach for automatic WordNet mapping using word sense disambiguation. The method has been applied to link Korean words from a bilingual dictionary to English WordNet synsets.

To clarify the description, an example is given. To link the first sense of Korean word "gwan-mog" to WordNet synset, we employ a bilingual Korean-English dictionary. The first sense of 'gwan-mog' has 'bush' as a translation in English and 'bush' has five synsets in WordNet. Therefore the first sense of 'gwan-mog' has five candidate synsets. Somehow we decide a synset {shrub, bush} among five candidate synsets and link the sense of 'gwan-mog' to this synset.

As seen from this example, when we link the senses of Korean words to WordNet synsets, there are semantic ambiguities. To remove the ambiguities we develop new word sense disambiguation heuristics and automatic mapping method to construct Korean WordNet based on the existing English WordNet.

This paper is organized as follows. In section 2, we describe multiple heuristics for word sense disambiguation for sense linking. In section 3, we explain the method of combination for these heuristics. Section 4 presents some experiment results, and section 5 will discuss some related researches. Finally we draw some conclusions and future researches in section 6. The automatic mapping-based Korean WordNet plays a natural Korean-English bilingual thesaurus, so it will be directly applied to Korean-English cross-lingual information retrieval as well as Korean monolingual information retrieval.

2 Multiple heuristics for word sense disambiguation

As the mapping method described in this paper has been developed for combining multiple individual solutions, each single heuristic must be seen as a container for some part of the linguistic knowledge needed to disambiguate the

* This research was supported by KOSEF special purpose basic research (1997.9 – 2000.8 #970-1020-301-3)

[†] Corresponding author

ambiguous WordNet synsets. Therefore, not a single heuristic is suitable to all Korean words collected from a bilingual dictionary. We will describe each individual WSD (word sense disambiguation) heuristic for Korean word mapping into corresponding English senses.

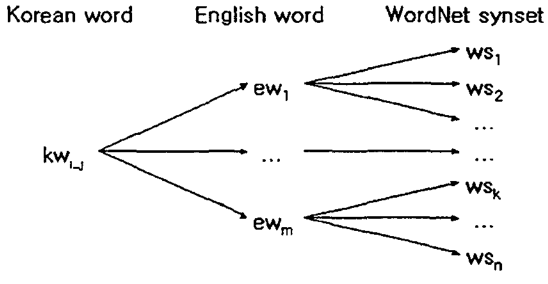

Figure 1: Word-to-Concept Association

Figure 1 shows the Korean word to WordNet synset association. The *j*-th sense of Korean word kw_i has *m* translations in English and *n* WordNet synsets as candidate senses. Each heuristic is applied to the candidate senses (ws_1, ... ,ws_n) and provides scores for them.

2.1 Heuristic 1: Maximum Similarity

This heuristic comes from our previous Korean WSD research (Lee and Lee, 2000) and assumes that all the translations in English for the same Korean word sense are semantically similar. So this heuristic provides the maximum score to the sense that is most similar to the senses of the other translations. This heuristic is applied when the number of translations for the same Korean word sense is more than 1. The following formula explains the idea.

$$H_1(s_i) = \max_{ew \in EW_i} \frac{1}{(n-1)+\alpha} \cdot \left(\sum_{j=1}^{n} support(s_i, ew_j) \ -1 \right)$$
$$\text{where } EW_i = \{ew \mid s_i \in synset(ew)\}$$

In this formula, $H_1(s_i)$ is a heuristic score of synset s_i, s_i is a candidate synset, *ew* is a translation into English, *n* is the number of translations and *synset(ew)* is the set of synsets of the translation *ew*. So Ew_i becomes the set of translations which have the synset s_i. The parameter α controls the relative contribution of candidate synsets in different number of translations: as the value of α increases, the candidate synsets in smaller number of translations get relatively less weight (α=0.5 was tuned experimentally). $support(s_i,ew)$ calculates the maximum similarity with the synset s_i and the translation *ew*, which is defined as :

$$support(s_i, ew) = \max_{s \in synset(ew)} S(s_i, s)$$

$$S(s_1, s_2) = \begin{cases} sim(s_1, s_2) & \text{if } sim(s_1, s_2) \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

Similarity measures lower than a threshold θ are considered to be noise and are ignored. In our experiments, θ=0.3 was used. $sim(s_1,s_2)$ computes the conceptual similarity between concepts s_1 and s_2 as in the following formula :

$$sim(s_1, s_2) = \frac{2 \times level(MSCA(s_1, s_2))}{level(s_1) + level(s_2)}$$

where $MSCA(s_1,s_2)$ represents the most specific common ancestor of concepts s_1 and s_2 and *level(s)* refers to the depth of concept s from the root node in the WordNet[1].

2.2 Heuristic 2: Prior Probability

This heuristic provides prior probability to each sense of a single translation as score. Therefore we will give maximum score to the synset of a monosemous translation, that is, the translation which has only one corresponding synset. The following formula explains the idea.

$$H_2(s_i) = \max_{ew \in EW_i} P(s_i \mid ew)$$
$$\text{where } EW_i = \{ew \mid s_i \in synset(ew)\}$$
$$P(s_i \mid ew_j) \approx \frac{1}{n_j}$$
$$\text{where } s_i \in synset(ew_j),\ n_j = |synset(ew_j)|$$

In this formula, n_j is the number of synsets of the translation ew_j.

2.3 Heuristic 3: Sense Ordering

(Gale et al., 1992) reports that word sense disambiguation would be at least 75% correct if a system assigns the most frequently occurring sense. (Miller et al., 1994) found that automatic

[1] We use English WordNet version 1.6

assignment of polysemous words in Brown Corpus to senses in WordNet was 58% correct with a heuristic of most frequently occurring sense. We adopt these previous results to develop sense ordering heuristic.

The sense ordering heuristic provides the maximum score to the most frequently used sense of a translation. The following formula explains the heuristic.

$$H_3(s_i) = \max_{ew \in EW_i} SO(s_i, ew)$$

$$\text{where } EW_i = \{ew \mid s_i \in synset(ew)\}$$

$$SO(s_i, ew) = \frac{\alpha}{x^\beta}$$

$$\text{where } s_i \in synset(ew)$$

$$\wedge\ synset(ew) \text{ is sorted by frequency}$$

$$\wedge\ s_i \text{ is the } x\text{-th synset in } synset(ew)$$

In this formula, x refers to the sense order of s_i in *synset*(*ew*): x is 1 when s_i is the most frequently used sense of *ew*. The information about the sense order of synsets of an English word was extracted from the WordNet.

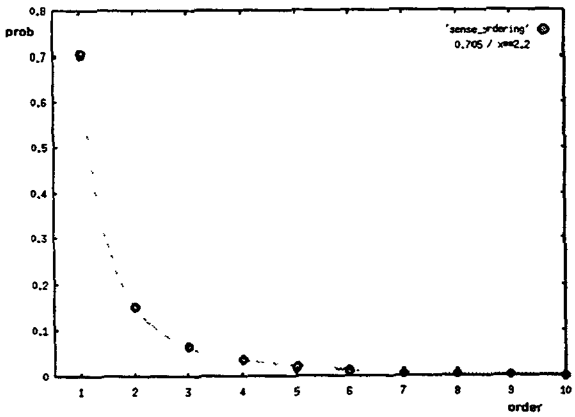

Figure 2: Sense distribution in SemCor

The value α=0.705 and β=2.2 was acquired from a regression of Figure 2 semcor corpus[2] data distribution.

2.4 Heuristic 4: IS-A relation

This heuristic is based on the following facts:

If two Korean words have an IS-A relation, their translations in English should also have an IS-A relation.

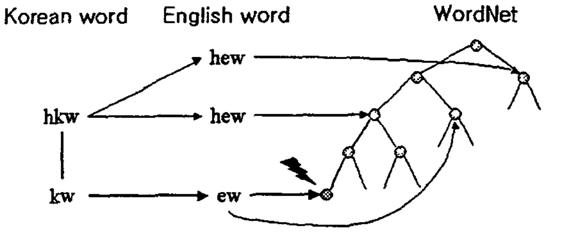

Figure 3: IS-A relation

Figure 3 explains IS-A relation heuristic. In figure 3, *hkw* is a hypemym of a Korean word *kw* and *hew* is a translation of *hkw* and *ew* is a translation of *kw*.

This heuristic assigns score 1 to the synsets which satisfy the above assumption according to the following formula:

$$H_4(s_i) = \max_{ew \in EW_i} IR(s_i, ew)$$

$$\text{where } EW_i = \{ew \mid s_i \in synset(ew)\}$$

$$IR(s_i, ew) = \begin{cases} 1 & \text{if } IsA(s_i, s_j) \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } s_i \in synset(ew), s_j \in synset(hew)$$

In this formula, $IsA(s_1,s_2)$ returns true if s_1 is a kind of s_2.

2.5 Heuristic 5: Word Match

This heuristic assumes that related concepts will be expressed using the same content words. Given two definitions – that of the bilingual dictionary and that of the WordNet – this heuristic computes the total amount of shared content words.

$$H_5(s_i) = \max_{ew \in EW_i} WM(s_i, ew)$$

$$\text{where } EW_i = \{ew \mid s_i \in synset(ew)\}$$

$$WM(s_i, ew) = sim(X, Y_i)$$

$$sim(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

In this formula, X is the set of content words in English examples of bilingual dictionary and Y_i is

[2] semcor is a sense tagged corpus from part of Brown corpus.

the set of content words of definition and example of the synset s_i in WordNet.

2.6 Heuristic 6: Cooccurrence

This heuristic uses cooccurrence measure acquired from the sense tagged Korean definition sentences of bilingual dictionary. To build sense tagged corpus, we use the definition sentences which have monosemous translation in bilingual dictionary. And we uses the 25 semantic tags of WordNet as sense tag :

$$H_6(s_i) = \max_{x \in Def} p(t_i \mid x)$$

$$\text{with } p = \hat{p} - Z_{(1-\alpha)/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\hat{p}(t_i \mid x) = \frac{Freq(t_i, x)}{Freq(x)}$$

In this formula, *Def* is the set of content words of a Korean definition sentence, t_i is a semantic tag corresponding to the synset s_i and n refers to *Freq(x)*.

3 Combining heuristics with decision tree learning

Given a candidate synset of a translation and 6 individual heuristic scores, our goal is to use all these 6 scores in combination to classify the synset as linking or discarding.

The combination of heuristics is performed by decision tree learning for non-linear relationship. Each internal node of a decision tree is a choice point, dividing an individual method into ranges of possible values. Each leaf node is labeled with a classification (linking or discarding). The most popular method of decision tree induction, employed here, is C4.5 (Quinlan, 1993).

Figure 4 shows a training phase in decision tree based combination method. In the training phase, the candidate synset ws_k of a Korean word is manually classified as linking or discarding and get assigned scores by each heuristic. A training data set is constructed by these scores and manual classification. The training data set is used to optimize a model for combining heuristics.

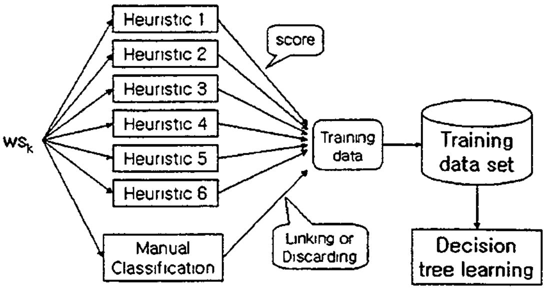

Figure 4: Training phase

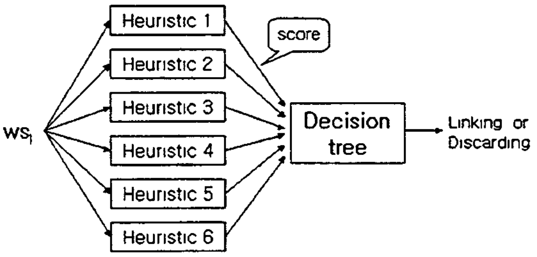

Figure 5: Mapping phase

Figure 5 shows a mapping phase. In the mapping phase, the new candidate synset ws_l of a Korean word is rated using 6 heuristics, and then the decision tree, which is learned in the training phase, classifies ws_l as linking or discarding. The synset classified as linking is linked to the corresponding Korean word.

4 Evaluation

In this section, we evaluate the performance of each six heuristics as well as the combination method. To evaluate the performance of WordNet mapping, the candidate synsets of 3260 senses of Korean words in bilingual dictionary was manually classified as linking or discarding.

We define 'precision' as the proportion of correctly linked senses of Korean words to all the linked senses of Korean words in a test set. We also define 'coverage' as the proportion of linked senses of Korean words to all the senses of Korean words in a test set.

Table 1 contains the results for each heuristic evaluated individually against the manually classified data. The test set here consists of the 3260 manually classified senses.

In general, the results of each heuristic seem to be poor, but are always better than the random choice baseline. The best heuristic according to the precision is the maximum similarity heuristic. But it was applied to only 59.51% of 3260 senses of Korean words. The results of each heuristic are better than the random mapping, with a statistically significance at the 99% level.

	Precision(%)	Coverage(%)
Random mapping	49.85	100.0
Heuristic 1	75.21	59.51
Heuristic 2	74.66	100.0
Heuristic 3	71.87	100.0
Heuristic 4	55.49	29.36
Heuristic 5	56.48	63.01
Heuristic 6	67.24	64.14

Table 1: Individual heuristics performance

	Precision(%)	Coverage(%)
Summing	84.61	100.0
Logistic regression	86.41	100.0
Decisioin tree	93.59	77.12

Table 2: Performance and comparison of the decision tree based combination

We performed 10-fold cross validation to evaluate the performance of the combination of all the heuristics using the decision tree – we split the data into ten parts, reserved one part as a validation set, trained the decision tree on the other nine parts and then evaluate the reserved part. This process is repeated nine times using each of the other nine parts as a validation set.

Table 2 shows the results of the other trials of the combination of all the heuristics. Summing is a way to simply sum all the scores of each heuristic. Then the candidate synset which has the highest summation of the scores is selected. Logistic regression, as described in (Hosmer and Lemeshow, 1989), is a popular technique for binary classification. This technique applies an inverse logit function and employs the iterative reweighted least squares algorithm. This technique determines the weight of each heuristic.

With the combination of the heuristics using summing, we obtained an improvement over maximum similarity heuristic (heuristic 1) of 9%, maintaining a coverage 100%. The decision tree is able to correctly map 93.59% of the senses of Korean words in bilingual dictionary, maintaining a coverage 77.12%.

Applying the decision tree to combine all the heuristics for all Korean words in bilingual dictionary, we obtain a preliminary version of the Korean WordNet containing 21654 senses of 17696 Korean nouns with an accuracy of 93.59% (±0.84% with 99% confidence).

5 Related works

Several attempts have been performed to automatically produce multilingual ontologies. (Knight & Luk 1994) focuses on the construction of Sensus, a large knowledge base for supporting the Pangloss Machine Translation system, merging ontologies (ONTOS and UpperModel) and WordNet with monolingual and bilingual dictionaries. (Okumura & Hovy 1994) describes a semi-automatic method for associating a Japanese lexicon to an ontology using a Japanese/English bilingual dictionary as a 'bridge'. Several lexical resources and techniques are combined in (Atserias et al., 1997) to map Spanish words from a bilingual dictionary to WordNet. In (Farreres et al., 1998), use of a taxonomic structure derived from a monolingual MRD is proposed as an aid to the mapping process.

This research is contrasted that it utilized bilingual dictionary to build monolingual thesaurus based on the existing popular lexical resources and used the combination of multiple unsupervided WSD heuristics.

6 Conclusion

This paper has explored the automatic construction of a Korean WordNet from pre-existing lexical resources – English wordNet and Korean/English bilingual dictionary. We presented several techniques for word sense disambiguation and their application to disambiguate the translations in bilingual dictionary. We obtained a preliminary version of the Korean WordNet containing 21654 senses of 17696 Korean nouns. In a series of experiments, we observed that the accuracy of mapping is over 90%.

References

Atserias J., Climent S., Farreras J., Rigau G. and Rodriguez H. (1997) Combining Multiple Methods for the Automatic Construction of

Multilingual WordNets. In *proceeding of the Conference on Recent Advances on NLP.*

Farreres X., Rigau G., and Rodriguez H.. (1998) Using WordNet for building WordNets. In *Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems.*

Gale W., Church K., and Yarowsky D. (1992) Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceeding of 30th Annual Meeting of the Association for Computational Linguistics.*

Hosmer Jr. and Lemeshow S. (1989) *Applied Logistic Regression.* Wiley, New York.

Knight K. and Luk S. (1994) Building a large-scale knowledge base for machine translation. In *Proceeding of the American Association for Artificial Intelligence.*

Miller G. (1990) Five papers on WordNet. *Special Issue of International Journal of Lexicography.*

Miller G., Chodorow M., Landes S., Leacock C. and Thomas R.. (1994) Using a semantic concordance for sense identification. In *Proceedings of the Human Language Technology Workshop.*

Okumura A. and Hovy E. (1994) Building Japanese-English Dictionary based on Ontology for Machine Translation. In *Proceedings of ARPA Workshop on Human Language Technology.*

Quinlan R.. (1993) *C4.5: Programs For Machine Learning.* Morgan Kaufmann Publishers.

Seungwoo Lee and Geunbae Lee. (2000) Unsupervised Noun Sense Disambiguation Using Local Context and Co-occurrence. In *Journal of Korean Information Science Society.* (in press)