Translation Selection through Source Word Sense Disambiguation and Target Word Selection

Hyun Ah Lee[†‡] and **Gil Chang Kim**[†]

[†]Dept. of EECS, Korea Advanced Institute of Science and Technology (KAIST),
373-1 Kusong-Dong Yusong-Gu Taejon 305-701 Republic of Korea
[‡]Daumsoft Inc., Daechi-Dong 946-12 Kangnam-Gu Seoul 135-280 Republic of Korea
halee@csone.kaist.ac.kr, gckim@cs.kaist.ac.kr

Abstract

A word has many senses, and each sense can be mapped into many target words. Therefore, to select the appropriate translation with a correct sense, the sense of a source word should be disambiguated before selecting a target word. Based on this observation, we propose a hybrid method for translation selection that combines disambiguation of a source word sense and selection of a target word. Knowledge for translation selection is extracted from a bilingual dictionary and target language corpora. Dividing translation selection into the two sub-problems, we can make knowledge acquisition straightforward and select more appropriate target words.

1 Introduction

In machine translation, translation selection is a process that selects an appropriate target language word corresponding to a word in a source language. Like other problems in natural language processing, knowledge acquisition is crucial for translation selection. Therefore, many researchers have endeavored to extract knowledge from existing resources.

As masses of language resources become available, statistical methods have been attempted for translation selection and shown practical results. Some of the approaches have used a bilingual corpus as a knowledge source based on the idea of Brown et al. (1990), but they are not preferred in general since a bilingual corpus is hard to come by. Though some latest approaches have exploited word co-occurrence that is extracted from a monolingual corpus in a target language (Dagan and Itai, 1994; Prescher et al., 2000; Koehn and Knight, 2001), those methods often fail in selecting appropriate words because they do not consider sense ambiguity of a target word.

In a bilingual dictionary, senses of a word are classified into several sense divisions and also its translations are grouped by each sense division. Therefore, when one looks up the translation of a word in a dictionary, she/he ought to resolve the sense of a word in a source language sentence, and then choose a target word among translations corresponding to the sense. In this paper, the fact that a word has many senses and each sense can be mapped into many target words (Lee et al., 1999) is referred to as the 'word-to-sense and sense-to-word' relationship, based on which we propose a hybrid method for translation selection.

In our method, translation selection is taken as the combined problem of sense disambiguation of a source language word and selection of a target language word. To disambiguate the sense of a source word, we employ both a dictionary based method and a target word co-occurrence based method. In order to select a target word, the co-occurrence based method is also used.

We introduce three measures for translation selection: sense preference, sense probability and word probability. In a bilingual dictionary, example sentences are listed for each sense division of a source word. The similarity between those examples and an input sentence can serve as a measure for sense disambiguation, which we define as sense preference. In the bilingual dictionary, target words are also recorded for each sense division. Since the set of those words can model each sense, we can calculate the probability of the sense by applying the co-occurrence based method to the set of words. We define the estimated probability as sense probability, which is taken for the other measure of sense disambiguation. Using co-occurrence, the probability of selecting a word from the set of trans-

lations can be calculated. We define it as word probability, which is a measure for word selection. Merging sense preference, sense probability and word probability, we compute preference for each translation, and then choose a target word with the highest translation preference as a translation.

2 Translation Selection based on 'word-to-sense and sense-to-word'

The 'word-to-sense and sense-to-word' relationship means that a word in a source language could have multiple senses, each of which might be mapped into various words in a target language. As shown in examples below, a Korean verb '*meok-da*' has many senses (work, deaf, eat, etc.). Also '*kkae-da*' has three senses (break, hatch, wake), among which the break sense of '*kkae-da*' is mapped into multiple words such as 'break', 'smash' and 'crack'. In that case, if
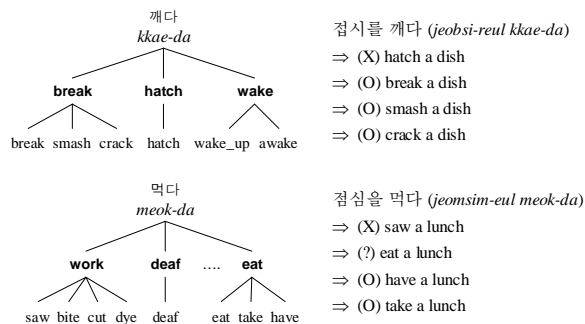
'hatch' is selected as a translation of '*kkae-da*' in a sentence '*jeobsi-reul kkae-da*', the translation must be wrong because the sense of '*kkae-da*' in the sentence is break rather than hatch. In contrast, any word of the break sense of '*kkae-da*' forms a well-translated sentence. However, selecting a correct sense does not always guarantee a correct translation. If the sense of '*meok-da*' in '*jeomsim-eul meok-da*' is correctly disambiguated as eat but an inappropriate target word is selected, an improper or unnatural translation will be produced like 'eat a lunch'.

Therefore, in order to get a correct translation, such a target word must be selected that has the right sense of a source word and forms a proper target language sentence. Previous approaches on translation selection usually try to translate a source word directly into a target word without considering the sense. Thus, they increase the complexity of the problem and suf-
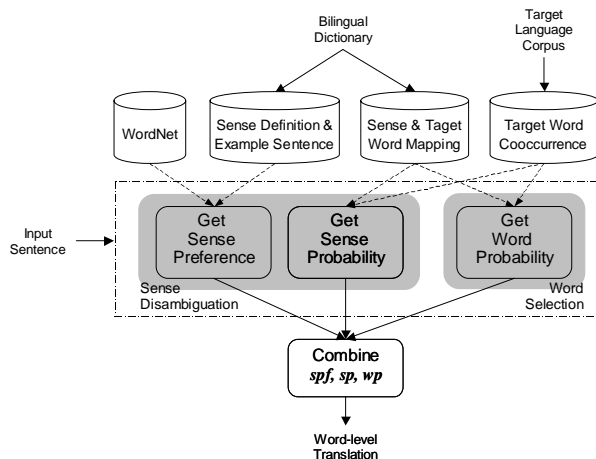

Figure 1: Our process of translation selection

fer from the problem of knowledge acquisition and incorrect selection of translation.

In this paper, we propose a method for translation selection that reflects 'word-to-sense and sense-to-word' relationship. We divide the problem of translation selection into sense disambiguation of a source word and selection of a target word. Figure 1 shows our process of selecting translation. We introduce three measures for translation selection: sense preference (spf) and sense probability (sp) for sense disambiguation, and word probability (wp) for word selection.

Figure 2 shows a part of an English-English-Korean Dictionary (EssEEK, 1995)[1]. As shown in the figure, example sentences and definition sentences are listed for each sense of a source word. In the figure, example sentences of the destroy sense of 'break' consist of words such as 'cup', 'piece', and 'thread', and those for the violate sense consist of words such as 'promise' and 'contract', which can function as indicators for the sense of 'break'. We calculate sense preference of a word for each sense by estimating similarity between words in an input sentence and words in example and definition sentences.

Each sense of a source word can be mapped into a set of translations. As shown in the example, the break sense of '*kkae-da*' is mapped into the set {'break', 'smash', 'crack'} and some words in the set can replace each other in a translated sentence like 'break/smash/crack a

---

[1]The English-English-Korean Dictionary used here is an English-Korean one, where English definitions are paired with Korean translations.

Figure 2: A part of an EEK Dictionary

dish'. Therefore, we can expect the behavior of those words to be alike or meaningfully related in a corpus. Based on this expectation, we compute sense probability by applying the target word co-occurrence based method to all words in the same sense division.

Word probability represents how likely a target word is selected from the set of words in the same sense division. In the example '*jeomsim-eul meok-da*', the sense of '*meok-da*' is eat, which is mapped into three words {'eat', 'have', 'take'}. Among them, 'have' forms a natural phrase while 'eat' makes an awkward phrase. We could judge the naturalness from the fact that 'have' and 'lunch' co-occurs in a corpus more frequently than 'eat' and 'lunch'. In other words, the level of naturalness or awkwardness of a phrase can be captured by word co-occurrence. Based on this observation, we use target word co-occurrence to get word probability.

3 Calculation of each measure

3.1 Sense Preference

Given a source word s and its k-th sense s^k, we calculate sense preference $spf(s^k)$ using the equation (1). In the equation, SNT is a set of all content words except s in an input sentence. DEF_{s^k} is a set of all content words in definition sentences of s^k, and EX_{s^k} is a set of all content words in example sentences of s^k, both of which are extracted from the dictionary.

$$spf(s^k) = \sum_{w_i \in SNT} \max_{w_d \in DEF_{s^k}} sim(w_i, w_d) \quad (1)$$
$$+ \sum_{w_i \in SNT} \max_{w_e \in EX_{s^k}} sim(w_i, w_e)$$

Sense preference is obtained by calculating similarity between words in SNT and words in DEF and EX. For all words in an input sentence ($w_i \in SNT$), we sum up the maximum similarity between w_i and all clue words (i.e. w_d and w_e). To get similarity between words ($sim(w_i, w_j)$), we use WordNet and the metric proposed in Rigau et al. (1997).

Senses in a dictionary are generally ordered by frequency. To reflect distribution of senses in common text, we use a weight factor $\phi(s^k)$ that is inversely proportional to the order of a sense s^k in a dictionary. Then, we calculate normalized sense preference to combine sense preference and sense probability.

$$spf_w(s^k) = \phi(s^k) \cdot \frac{spf(s^k)}{\sum_i spf(s^i)} \quad (2)$$

$$spf_{norm}(s^k) = \frac{spf_w(s^k)}{\sum_i spf_w(s^i)} \quad (3)$$

3.2 Sense Probability

Sense probability represents how likely target words with the same sense co-occur with translations of other words within the input sentence. Let us suppose that the i-th word in an input sentence is s_i and the k-th sense of s_i is s_i^k. Then, sense probability $sp(s_i^k)$ is computed as follows:

$$n(t_{iq}^k) = \sum_{(s_j, m, c) \in \Theta(s_i)} \sum_{p=1}^{m} \frac{f(t_{iq}^k, t_{jp}, c)}{f(t_{iq}^k) + f(t_{jp})} \quad (4)$$

$$sp(s_i^k) = \hat{p}_{sense}(s_i^k) = \frac{\sum_q n(t_{iq}^k)}{\sum_x \sum_y n(t_{iy}^x)} \quad (5)$$

In the equation, $\Theta(s_i)$ signifies a set of words that co-occur with s_i on syntactic relations. In an element (s_j, m, c) of $\Theta(s_i)$, s_j is a word that co-occurs with s_i in an input sentences, c is a syntactic relation[2] between s_i and s_j, and m is the number of translations of s_j. Provided that the set of translations of a sense s_i^k is T_i^k and a member of T_i^k is t_{iq}^k, the frequency of co-occurring t_{iq}^k and t_{jp} with a syntactic relation c is denoted as $f(t_{iq}^k, t_{jp}, c)$, which is extracted from target language corpora.[3] Therefore,

---

[2] We guess 37 syntactic relations in English sentences based on verb pattern information in the dictionary and the results of the memory based shallow parser (Daelemans et al., 1999): `subj-verb`, `obj-verb`, `modifier-noun`, `adverb-modifiee`, `etc`.

[3] When '*jeobsi*' has 4 translations ('plate', 'dish', 'saucer', 'platter'), Θ(*kkae-da*) has a member (*jeob-*

$n(t_{iq}^k)$ in the equation (4) represents how frequently t_{iq}^k co-occurs with translations of s_j. By summing up $n(t_{iq}^k)$ for all target words in T_i^k, we obtain sense probability of s_i^k.

3.3 Word Probability

Word probability represents how frequently a target word in a sense division co-occurs in a corpus with translations of other words within the input sentence. We denote word probability as $wp(t_{iq}^k)$ that is a probability of selecting t_{iq}^k from T_i^k. Using $n(t_{iq}^k)$ in the equation (4), we calculate word probability as follows:

$$wp(t_{iq}^k) = \hat{p}_{word}(t_{iq}^k) = \frac{n(t_{iq}^k)}{\sum_x n(t_{ix}^k)} \qquad (6)$$

3.4 Translation Preference

To select a target word among all translations of a source word, we compute translation preference (tpf) by merging sense preference, sense probability and word probability. We simply sum up the value of sense preference and sense probability as a measure of sense disambiguation, and then multiply it with word probability as follows:

$$tpf(t_{iq}^k) = (\delta \cdot spf_{norm}(s_i^k) + (1-\delta) \cdot sp(s_i^k)) \qquad (7)$$
$$\cdot \, wp(t_{iq}^k)/\eta(T_i^k)$$

In the equation, $\eta(T_i^k)$ is a normalizing factor for $wp(t_{iq}^k)$[4], and δ is a weighting factor for sense preference. We select a target word with the highest tpf as a translation.

When all of the $n(t_{iq}^k)$ in the equation (4) are 0, we use the following equation for smoothing, which uses only frequency of a target word.

$$n(t_{iq}^k) = \sigma \cdot \frac{f(t_{iq}^k)}{\sum_p f(t_{ip}^k)} \qquad (8)$$

---

si, 4, `verb-obj`) in the example of '*jeobsi-reul kkae-da*', and then the following frequencies will be used: f(break, plate, `verb-obj`), f(break, dish, `verb-obj`),..., f(crack, platter, `verb-obj`).

[4]Because $wp(t_{iq}^k)$ is a probability of t_{iq}^k within T_i^k, $wp(t_{iq}^k)$ becomes 1 when T_i^k has only one element. $\eta(T_i^k)$ is used to make the maximum value of $wp(t_{iq}^k)/\eta(T_i^k)$ to 1. Hence, $\eta(T_i^k) = \max_j wp(t_{ij}^k)$. Using $\eta(T_i^k)$, we can prevent discounting translation preference of a word, the sense of which has many corresponding target words.

4 Evaluation

We evaluated our method on English-to-Korean translation. EssEEK (1995) was used as a bilingual dictionary. We converted this English-English Korean dictionary into a machine readable form, which has about 43,000 entries, 34,000 unique words and 80,000 senses. From the dictionary, we extracted sense definitions, example sentences and translations for each sense. Co-occurrence of Korean was extracted from Yonsei Corpus, KAIST corpus and Sejong Corpus using the method proposed in Yoon (2001). The number of extracted co-occurrence is about 600,000.

To exclude any kind of human intervention during knowledge extraction and evaluation, we evaluated our method similarly with Koehn and Knight (2001). English-Korean aligned sentences were collected from novels, textbooks for high school students and sentences in a Korean-to-English bilingual dictionary. Among those sentences, we extracted sentence pairs in which words in the English sentence satisfy the following condition: if the paired Korean sentence contains a word that is defined in the dictionary as a translation of the English word. We randomly chose 945 sentences as an evaluation set, in which 3,081 words in English sentences satisfy the condition. Among them, 1,653 are nouns, 990 are verbs, 322 are adjectives, and 116 are adverbs.

We used Brill's tagger (Brill, 1995) and Memory-Based Shallow Parser (Daelemans et al., 1999) to analyze English sentences. To analyze Korean sentences, we used a POS tagger for Korean (Kim and Kim, 1996).

First, we evaluated the accuracies of sense preference (spf) and sense probability (sp). If any target word of the sense with the highest value is included in an aligned target language sentence, we regarded it as correct result.

The accuracy of sense preference is shown in Table 1. In the table, a w/o DEF row shows the result produced by removing the *DEF* clue in the equation (1), and a w/o ORDER row is the result produced without the weight factor $\phi(s^k)$ of the equation (2). As shown in the table, the result with the weight factor and without a sense definition sentence is best for all cases. We will discuss this result in the next section.

The accuracy of sense probability is shown in

Table 1: Accuracy of sense preference (spf)

		noun	verb	adj.	adv.	all
w/o ORDER	w/o DEF	59.23%	42.93%	45.03%	34.48%	52.47%
	with DEF	57.65%	40.20%	49.07%	33.62%	51.14%
with ORDER[5]	w/o DEF	**66.30%**	**48.48%**	**59.32%**	**44.83%**	**59.94%**
	with DEF	65.76%	45.56%	56.52%	42.24%	58.41%

Table 2: Accuracy of sense probability (sp)

	noun	verb	adj.	adv.	all
with all cooc word	**52.49%**	30.49%	38.94%	59.35%	**46.52%**
with case cooc word	49.62%	**31.33%**	**40.12%**	**60.98%**	45.41%

Table 3: Accuracy of translation selection

		noun	verb	adj.	adv.	all
random selection		11.11%	3.92%	7.41%	11.12%	6.77%
1st translation of 1st sense		12.89%	12.12%	7.45%	2.59%	11.86%
most frequent		46.34%	27.58%	31.99%	37.93%	38.68%
spf		53.72%	38.18%	39.75%	44.83%	46.98%
with all cooc word	sp	42.88%	19.28%	26.25%	39.02%	35.04%
	wp	51.66%	31.41%	32.92%	46.55%	43.10%
	$sp \times wp$	51.97%	31.92%	33.54%	50.00%	43.62%
	$spf \times wp$	**55.23%**	40.00%	41.30%	50.00%	**50.17%**
	$(spf+sp) \times wp$[6]	54.99%	34.55%	36.34%	50.86%	46.42%
with case cooc word	sp	38.90%	20.20%	25.96%	37.40%	33.05%
	wp	48.70%	34.04%	35.40%	45.69%	42.58%
	$sp \times wp$	49.12%	33.74%	36.34%	50.86%	43.01%
	$spf \times wp$	53.60%	**42.22%**	**42.86%**	50.86%	49.93%
	$(spf+sp) \times wp$[6]	51.18%	36.46%	38.20%	**53.45%**	45.28%

Table 2. In the equation (4), we proposed to use syntactic relations to get sense probability. In the table, the result obtained by using words on a syntactic relation (with case cooc word) is compared to that obtained by using all words within a sentence (with all cooc word). As shown, the accuracy for nouns is higher when using all words in a sentence, whereas the accuracy for others is higher when considering syntactic relations.

Table 3 shows the result of translation selection.[7] We set three types of baseline - random selection, most frequent, 1st translation of 1st sense[8]. We conducted the experiment al-

tering the combination of spf, sp and wp. An spf row shows the accuracy of selecting the first translation of the sense with the highest spf, and an sp row shows the accuracy of selecting the first translation of the sense with the highest sp. A wp row shows the accuracy of using only word probability. In other words, it is obtained through assuming all translations of a source word to have the same sense. Therefore, the result in a wp row can be regarded as a result produced by the method that uses only target word co-occurrence. An $sp \times wp$ row is the result obtained without spf_{norm}, and an $spf \times wp$ row is the result obtained without sp in the equation (8). An $(spf+sp) \times wp$ row is the result of combining all measures.

For the best case, the accuracy for nouns is 55.23%, that for verbs is 42.22% and that for adjectives is 42.86%. Although we use a simple method for sense disambiguation, our results of translation selection are superior to the results that are produced using only target word co-

---

[5]In the case of ϕ(1st sense)=1.5, ϕ(2nd sense)=1.3, ϕ(3rd sense)=1.15, and ϕ(remainder)=1

[6]In the case of δ=0.6, which shows the best result in the experiment.

[7]To get spf, we use only the best combination of clues - with the weight factor and without sense definition sentences.

[8]Result of selecting translation that is recorded at first place in the first sense division of a source word.

occurrence (*wp* rows).

5 Discussion

We could summarize the advantage of our method as follows:

- reduction of the problem complexity
- simplifying knowledge acquisition
- selection of more appropriate translation
- use of a mono-bilingual dictionary
- integration of syntactic relations
- guarantee of robust result

The figure below shows the average number of senses and translations for an English word in EssEEK (1995). The left half of the graph is for all English words in the dictionary, and the right half is for the words that are marked as frequent or significant words in the dictionary. On the average, a frequently appearing English word has 2.82 senses, each sense has 1.96 Korean translations, and consequently an English word has 5.55 Korean translations. Moreover, in our evaluation set, a word has 6.86 senses and a word has 14.78 translations. Most approaches on translation selection have tried to translate a source word directly into a target word, thus the complexity of their method grows proportional to the number of translations per word. In contrast, the complexity of our method is proportional to only the number of senses per word or the number of translations per sense. Therefore, we could reduce the complexity of translation selection.

When the degree of ambiguity of a source language is n and that of a target language is m,
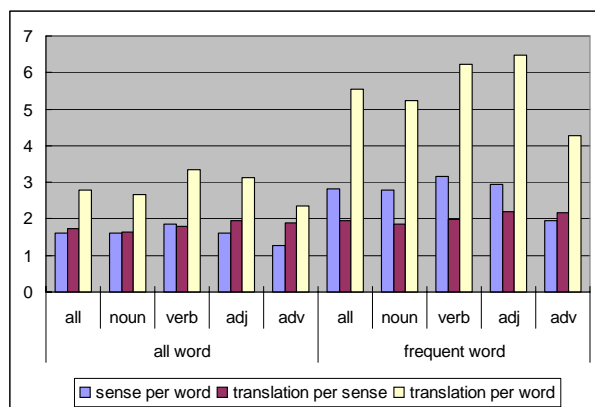

Figure 3: The average number of senses and translations for an English word

the complexity of translation is nearly $n \times m$. Therefore, knowledge acquisition for translation is more complicated than other problems of natural language processing. Although the complexity of knowledge acquisition of the methods based on a target language corpus is only m, ignorance of a source language results in selecting inappropriate target words. In contrast, by splitting translation selection into two sub-problems, our method uses only existing resources - a bilingual dictionary and a target language monolingual corpus. Therefore, we could alleviate the complexity of knowledge acquisition to around $n+m$.

As shown in the previous section, our method could select more appropriate translation than the method based on target word co-occurrence, although we used coarse knowledge for sense disambiguation. It is because the co-occurrence based method does not consider sense ambiguity of a target word. Consider the translation of the English phrase 'have a car'. A word 'car' is translated into a Korean word '*cha*', which has many meanings including car and tea. A word 'have' has many senses - posses, eat, drink, etc., among which the drink sense is mapped into a Korean verb '*masi-da*'. Because of the tea sense of '*cha*', the frequency of co-occuring '*cha*' and '*masi-da*' is dominant in a corpus over that of all other translations of 'car' and 'have'. In that case, the method that uses only word co-occurrence (*wp* in our experiment) translated 'have a car' into an incorrect sentence '*cha-reul masi-da*' that means 'have a tea'. In contrast, our method produced a correct translation '*cha-reul kaji-da*', because we disambiguate the sense of 'have' as posses by employing sense preference before using co-occurrence.

In the experiment, the combination $spf \times wp$ gets higher accuracy than $(spf+sp) \times wp$ for most cases. It is due to the low accuracy of sp, which is also caused by the ambiguity of a target word as shown in the example of 'have a car'. Nevertheless, we do not think it means sense probability is useless. The accuracies of $sp \times wp$ are almost better than those of wp, therefore sense probability can work as a countermeasure when no knowledge of a source language exists for disambiguating the sense of a word.

In this paper, we used a mono-bilingual dictionary, which has more information than a

common bilingual dictionary including sense definitions in a source language and syntactic patterns for the predicate. While previous research on sense disambiguation that exploits those kinds of information has shown reliable results, in our experiment, the use of sense definitions lowers the accuracy. It is likely because we used sense definitions too primitively. Therefore, we expect that we could increase the accuracy by properly utilizing additional information in the dictionary.

Many studies on translation selection have concerned with translation of only nouns. In particular, some of them use all co-occurring words within a sentence, and others use a few restricted types of syntactic relations. Even more, each syntactic relation is utilized independently. In this paper, we proposed a statistical model that integrates various syntactic relations, with which the accuracies for verbs, adjectives, and adverbs increase. Nevertheless, since the accuracy for nouns is higher when using all words in a sentence, it seems that syntactic relations may be differently used for each case.

The other advantage of our method is robustness. Since our method disambiguates the sense of a source language word using reliable knowledge in a dictionary, we could avoid selecting or generating a target word the meaning of which is definitely improper in given sentences like 'hatch a dish' for 'jeobsi-reul kkae-da' and 'cha-reul masi-da' for 'have a car'.

6 Conclusion

In this paper, we proposed a hybrid method for translation selection. By dividing translation selection into sense disambiguation of a source word and selection of a target word, we could simplify knowledge acquisition and select more appropriate translation.

Acknowledgement

References

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing : A case study in part of speech tagging. *Computational Linguistics*, 21(4).

Peter F. Brown, John Cocke, Vincent Della Pietra, Stephen Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2).

Walter Daelemans, Sabine Buchholz, and Jorn Veenstra. 1999. Memory-based shallow parsing. In *Proceedings of the 3rd International Workshop on Computational Natural Language Learning (CoNLL-'99)*.

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monoligual corpus. *Computational Linguistics*, 20(4).

EssEEK, 1995. *Essence English-English Korean Dictionary*. MinJung SeoRim.

Jae-Hoon Kim and Gil Chang Kim. 1996. Fuzzy network model for part-of-speech tagging under small training data. *Natural Language Engineering*, 2(2).

Philipp Koehn and Kevin Knight. 2001. Knowledge sources for word-level translation models. In *Proceedings of Empirical Methods in Natural Language Processing conference (EMNLP-2001)*.

Hyun Ah Lee, Jong C. Park, and Gil Chang Kim. 1999. Lexical selection with a target language monolingual corpus and an mrd. In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-'99)*.

Detlef Prescher, Stefan Riezler, and Mats Rooth. 2000. Using a probabilistic class-based lexicon for lexical ambiguity resolution. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*.

German Rigau, Jordi Atserias, and Eneko Agirre. 1997. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-'97)*.

Juntae Yoon. 2001. Efficient dependency analysis for korean sentences based on lexical association and multi-layered chunking. *Literary and Linguistic Computing*, 16(3).