Chinese-Korean Word Alignment Based on Linguistic Comparison

Jin-Xia Huang[*,**]
[*]KORTERM, AITrc, Computer Science Department, Korea Advanced Institute of Science and Technology
373-1 Yusong-gu, Gusong-dong, Daejon 305-701, Republic of Korea
[**]Yanbian University of Science & Technology
Yanji City, Jilin Province, 133001, P.R.China
hgh@world.kaist.ac.kr

Key-Sun Choi
KORTERM, AITrc, Computer Science Department, Korea Advanced Institute of Science and Technology
373-1 Yusong-gu, Gusong-dong, Daejon 305-701, Republic of Korea
kschoi@world.kaist.ac.kr

Abstract

Word alignment problem between parallel corpora is based on the similar characteristics of two aligned words in two languages. We investigate linguistic-knowledge-based word similarity measures while other previous works heavily rely on statistical information, and their limits will be discussed. Linguistic knowledge is acquired from linguistic comparison of all layers between two languages, for Chinese and Korean in this paper.

1 Introduction

1.1 Previous works

The bilingual corpus provides more information than the monolingual one (Dagan, 1991). In recent years, much works have been done on word alignment after the research on section, paragraph, sentence, and phrase level alignment. Word alignment works for the automatic construction of bilingual dictionaries, bilingual patterns and other useful resources, and further, it works for the various applications such as machine translation and word sense disambiguation.

Statistical approach has been used as a main technique in most alignment systems (Gale, 1991; Brown, 1993; Dagan, 1994; Kay, 1993; Wu, 1994; Smadja, 1996; Tanaka, 1999). Correlation information is mainly employed in statistical approach, and other similarities like character, word length and position are employed in it (Dagan, 1994; Fung, 1998; Kevin, 1999). Clues for alignment are investigated, for example, functional word and context seed word in (Brown, 1993; Shin & Choi, 1996).

In contrast to the systems that mainly rely on statistical approach, Ker (1997) uses a class-based algorithm without any statistical technique in the English-Chinese word alignment. Ker's approach is claimed to overcome the lower coverage of statistical approaches while gaining high precision. Her work shows us the feasibility of pure linguistic approach to enhance the resolution of alignment problem.

1.2 Problem Definition

In the most previous studies, the “alignment” was usually defined by "aligning word (or text, phrase, section, etc.) to its translation" (Gale, 1991; Shin & Choi, 1996). It seems that the concept of alignment is so obvious that no one has concerned for the problem, "what is the translation of an original word". In this paper, we would like to clarify the definition as follows:

> "Alignment" is to find out the translated version of the given source language. "Word/phrase alignment", therefore, is to find out, from the aligned pairs of sentences, two words/phrases with the highest semantic similarities and the highest syntactic similarities.

Based on this definition given above, we can easily point out that alignment problem is essentially the problem of bilingual word sense and syntactic similarity.

Traditional statistical approaches have been testified to be effective in the resolution of alignment problem: statistical information reflects word similarity in some stages. But this approach gets good result mainly in the alignment of between the languages that belong

to the same language family (Brown, 1993; Dagan, 1994; Dan, 1997), and shows limitation in coverage even after training with extremely large bilingual corpus (Ker, 1997). To the languages that do not belong to the same language family, statistical approaches have shown limited coverage and low precision, even after the employment of additional information (Shin and Choi, 1996; Turcato, 1998). This result is not surprising because statistical approach is just an indirect way to obtain the word similarity.

Then, what is the more direct information in getting bilingual word similarity? In monolingual processing, some resources such as dictionary, thesaurus and WordNet have been used customarily. Alignment needs bilingual information, so we attempt to use bilingual dictionary instead of the monolingual dictionary. Thesaurus and WordNet have almost never been used in bilingual alignment except Ker (1997) because they normally contain only monolingual information, but Ker shows us a sound approach to make use of monolingual thesaurus in bilingual alignment.

Though there are many differences between some Asian language pairs like Chinese and Korean (or Japanese and Chinese, or Japanese and Korean), but we know that there are also many similarities between them. And so, as the result, we believe that linguistic knowledge will be more close to the bilingual word sense or syntactic similarity than some statistical information or only word position in sentence.

2 Linguistic comparison between Chinese and Korean

Korean language belongs to Altai language family while Chinese language belongs to Sino-Tibetan, so it is not surprising that there are many differences between them. To get useful information for the word alignment, we will focus mainly on their common points more than their differences. We will compare the linguistic properties of Chinese and Korean from the three viewpoints - character, lexicon, and syntax.

2.1 Character Comparison

Differently from the language pairs such as French-English or some Europe language pairs, Chinese-Korean characters look very different from each other, because Chinese characters are ideograph while Korean characters are phonogram. But actually, because Korean characters are phonogram, and there is historical relation between Chinese and Korean, almost all of the Chinese characters can be converted to one or several Korean characters, these Korean characters express the pronunciation of the Chinese characters in Korean (e.g., for '名' in Chinese, which pronunciation is *'ming'*, it can be unically converted to Korean character '명' pronounced by *'myung'*).

We have constructed a Chinese-Korean Character Transfer Table (CKCT Table) to reflect the correspondence relation between them. The 6763 Chinese characters that are listed in GB2313-80 Chinese standard code table can be converted to 436 Korean characters for their Korean localized pronuncation.

2.2 Word Comparison

We can try to find the lexical similarities between Chinese and Korean languages from three aspects: word formation, part-of-speech (POS) and lexical internal structure.

2.2.1 Word Formation

About 60% of the Korean colloquial words are derived from Chinese words (Chinese-Korean words) (Choi, 1989). Normally, these Korean words have similar forms with the related Chinese words when transferring the Chinese words to their Korean pronunciation. (e.g.,"[C]和平 (*heping*) → [CK](*hwapyeong*) ⇔ [K]평화(*pyeonghwa*)", "[C]办公室(*bangongshi*) → [CK](*pangongsil*) ⇔ [K]사무실(*samusil*)), where [C], [CK] and [K] stand for Chinese word, Korean pronuncation of Chinese, and its corresponding Korean word, respectively.

It seems that word formation similarity between Chinese and Korean gives good word alignment. But because of the long history of Chinese character use in Korea, the word formation and their concept are quite different. For example, *zhailu* (摘录) in Chinese stands for “excerpt” in English, which is expressed by *balcwi* (拔萃) in Korean, but *bacui* (拔萃) in contemporary Chinese normally stands for "supereminence" in English. Besides this, there is noise when using word formation similarity in alignment, for example, "청실(*ceungsil*)" (means "blue thread") contains a similar word formation with “교실

(*gyosil*)" (教室; "classroom") while their meanings are very different from each other. Such noise will be more serious for the "short" words that are composed of only one or two characters. For example, "*sil*" in Korean is mapped into many different Chinese characters like "室, 失, 实" in corresponding Chinese pronunciation "*shi*".

Similarity of the lexical formation is on the prolongation of the character similarity. Similarity of lexical formation exists between other language pairs also, and this property has already been used in the other previous works (Church, 1993).

2.2.2 Part-of-Speech (POS)

POS similarity indicates the regularity between the POS of the source word and its translation. For example, if a word is a pronoun in source language, then it is highly probable that the translation of the word also belongs to a pronoun. POS similarity attracted much attention by computational linguistic researchers long before, and has been made use in several previous alignment systems (Dagan, 1994; Shin and Choi, 1996).

There is POS similarity between Chinese and Korean also. Our experiment shows that about 77.1% Chinese nouns are mapped to Korean noun (common noun), while only less than 8% of them are translated to Korean verb. But it is not always so optimistic, for example, only 34.4% of Chinese verbs are translated into Korean verb, while 35.1% of them are translated to Korean noun. It is the reason why we try to use more kinds of linguistic knowledge in our alignment study.

2.2.3 Lexical Internal Structure

Different from other language words, Chinese words have internal structure. For example, the verb " 下雨(*xiayu*) (rain)" is composed of two words, one is verb "下(*xia*)(fall,drop…)" and the other one is noun "雨(*yu*)(rain)", therefore, the internal structure of the word " 下雨(*xiayu*)(rain)" is "verb+noun". We call it "phrasal word" because there is a phrasal structure inside of word.

We found that in most cases of 1:n (where n is the number of corresponding words) correspondence, Chinese words have some specific POS and internal structures. Chinese phrasal words and their corresponding Korean phrases hold similar syntactic structure. We name it "lexical internal structure similarity". For example, consider a fragment "it rains":

[C] *xia+yu* (下雨)⇔[K] *bi/ga+o/da* (비가 오다)
verb+noun ⇔ noun/SUBJ+verb/ending
([E] *come down+rain*⇔ *rain+come down*)

The lexical structure transfer rules (*e.g.,* "[C] verb+noun ⇔ [K]noun/CASE +verb/ending") can be constructed semi-automatically.

2.3 Syntactic Comparison

Word position (e.g., between English and French), functional word (e.g., between Korean and English) or POS information reflects syntactic information. They have been used with statistical approach in many previous works (Gale, 1991; Brown 1993; Shin & Choi, 1996). But in the alignment of Chinese and Korean, word positions in a sentence are not synchronized enough because their word orders are quite different. Chinese is SVO type language while Korean is SOV one, and both of their word orders are quite flexible. Additionally, Chinese word order is reflected more by semantic element than by syntactic one (Li, 1981).

But it does not mean that there is no syntactic similarity between Chinese and Korean. Syntactic similarity indicates that there is syntactic regularity in syntactic structure transformation. This property can be described in simple transfer patterns that contain no embedded structure (as in right hand side of next examples).

"[C]新/adj 书/noun(*xin shu*)" ⇔ "[K] 새/adj 책/noun (*sae caeg*)" ([C]adj noun ⇔ [K]adj noun)
"[C]讨论/v 讨论/v (*taolun taolun*)" ⇔ "[K]의논/noun+하여 보자 (*yinonha'yeo boja*)" ([C]verb1 verb1 ⇔ [K]noun+하여 보자)

In Chinese-Korean alignment, using word position in simple transfer pattern is more accurate than only using word position in sentence. And the precision of simple transfer pattern is higher than of POS information, because POS information reflects only the information of one word without context, while the simple transfer pattern reflects the context information of the word.

In practice, we can employ probabilistic information instead of transfer pattern. We will describe it in the session 3.3.

3 Chinese-Korean Word Alignment

3.1 Alignment Object

Alignment objects are restricted to some substantive (content words) in both Chinese and Korean. The standard of the exclusion is that, if most words of one POS have one to zero (1:0) correspondence relation in alignment, it will be excluded. As a result, all of the expletives and quantifier of Chinese, and all of the function words of Korean are excluded.

3.2 Resource and Information Used in the System

Table 1 shows the linguistic resources we use in our system and information they can provide.

Resource	Information
CKCT Table	Similarity of lexical formation
Bilingual Dictionary	Semantic similarity and lexical internal structure information
Bilingual ClassNet	Conceptual Similarity
Simple Pattern with Probability	Syntactic Similarity
Bilingual Corpus	Correlation information of bilingual words

Table 1. Resource and information used in the system

Bilingual dictionary provides us the target words that have the highest semantic similarity with the source words, and it is useful in obtaining the lexical structure similarity. Bilingual ClassNet is constructed with Korean and Chinese monolingual thesauri, and it provides us the conceptual similarity between Chinese and Korean words. CKCT Table helps us get similarity of the word formation. Simple pattern database with probabilistic information will be helpful in getting syntactic similarity, and we reflect the lexical internal structure similarity with simple pattern also. Finally, bilingual corpus provides us correlation information of bilingual words as it is known to us.

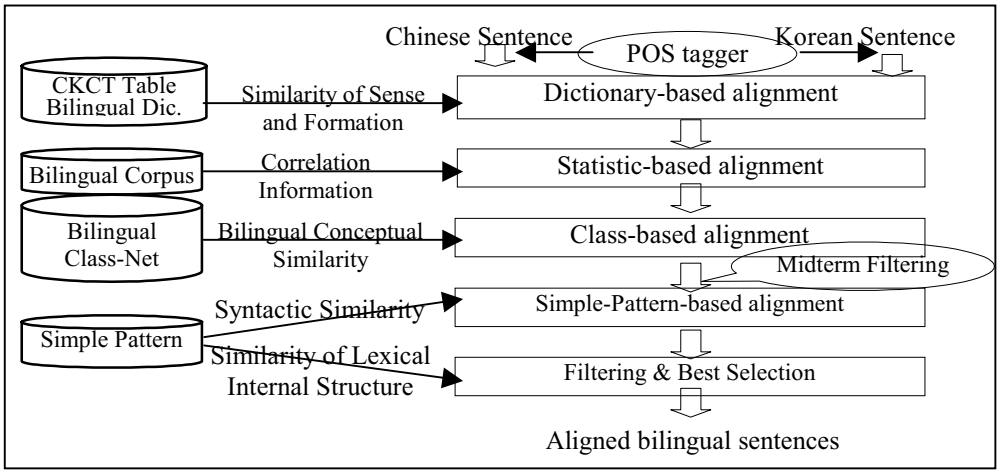

Figure 1. Chinese-Korean alignment system architecture

3.3 Alignment Algorithm

We use Chinese and Korean POS tagger as preprocessor of our system. Figure 1 is the Chinese-Korean alignment system architecture.

3.3.1 Dictionary-Based Alignment

We use CKCT and bilingual dictionary in this stage, and employed dice coefficient equation (Dice, 1945) to measure the similarity of lexical formation. The algorithm is as follows:

1. Using CKCT table, transferfing the given Chinese word c_i to its Korean character string by converting characters one by one, add it to empty set k_{ci}.
2. Search the Korean translations of c_i from bilingual dictionary. Add them to the set k_{ci}. We consider one Korean phrase that contains no space as a word. Delete postfix of the Korean word when it is verb, adverb or adjective.
3. Calculate similarity of c_i and k_j - $WordSim(c_i,k_j)$ with equation (1). Only the result $WordSim(c_i, k_j)$> t_1(t_1 is threshold) will remain.

$$WordSim(c,k) = d \times \max_{k_{c_i} \in k_c} \frac{2 \times |k_{c_i} \cap k|}{|k_{c_i}| + |k|} \quad (1)$$

Where

c = Chinese word of given sentence
k_c = Korean lexical set corresponding to c
k_{Ci} = i th element of set k_c
k = Korean morpheme of given sentence
$|k|$ = Total number of the characters in k
$|k_{Ci}|$ = Total number of the characters in ith element of set k_C
d : A constant. (If k_{Ci} and k have diffirent POS tagger, or $|c|$=1 and $|k|$=1, d<1.00; otherwise d = 1.00)

3.3.2 Statistic-Based Alignment

We will align high co-occurred words that share low formation similarity or can not be found in bilingual dictionary. T-score was used by Fung (1996) as a confidence measure after using MI information. Because it does not favour rare words as much as the MI does, and for time saving[1], we choose only t-score in our system. T-score reflects correlation of bilingual words as equation (2). Only the result that is bigger than threshold will remain.

$$t-score(c,k) = \frac{N_{ck} \times Total - N_c \times N_k}{Total \times \sqrt{Total}} \quad (2)$$

Where

c = Chinese word of given sentence
k = Korean morpheme of given sentence
N_c = Occurrence times of c in corpus
N_k = Occurrence times of k in corpus
N_{ck} = Co-occurrence times of c and k
$Total$ = Sentence pair number of corpus

3.3.3 Class-Based Alignment

We constructed Bilingual ClassNet with Chinese Tongyici Cilin (Synonym Forest, Mei and et.al, 1983) and Korean Thesaurus automatically (Huang & Choi, 1999), and employed it in this stage. Here are examples of the ClassNet:

ClassSim([C]Ab01,[K]12040)[2] = 0.548;
(Ab01:man,woman;12040: man and woman)
ClassSim([C]Ab01,[K]12050) = 0.514;
(Ab01:man,woman;12050:oldie and younger)
ClassSim([C]Ab01,[K]12100) = 0.324;
(Ab01: man, woman; 12100: family)
… …

$$WordSim(c_i,k_j) = \underset{p=1,q=1}{\overset{n,m}{MAX}}(ClassSim(C_{ip},K_{jq})),$$
$$where\ \ c_i \in \{C_{i1},C_{i2},...,C_{in,}\},\ \ k_j \in \{K_{j1},K_{j2},...,K_{jm}\} \quad (3)$$

Search all of the classes $\{C_{i1,}\ C_{i2,}\ ...\ C_{in}\}$ of the given word c_i from Chinese thesarus Cilin, and all of the classes $\{K_{j1,}\ K_{j\ 2,}\ ...\ K_{jm}\}$ that the Korean word k_j belongs to. Get the word similarity of c_i and k_j by equation (3). If the similarity is bigger than threshold t_2, remain it.

3.3.4 Midterm Filtering

To raise the precision, system will filter the alignment result before stage 4 with heuristics. In this stage, all of the alignment candidates will be marked with different level by their similarity that have been gotten in dictionary-based, statistical-based and class-based alignment. If there are more than two alignment candidates to one Chinese word, the best one will be selected. If the similarities are very close, then they will be all remained.

3.3.5 Simple-Pattern-Based Alignment

In this stage, we will align the words that have failed to be aligned yet, using simple pattern that with probabilistic information. Aligned word information and word position information will be employed, and only the word position inside the range of simple pattern will be considered useful. Let's look at the algorithm of this stage with an example.

Assume that in a given Chinese sentence, there is a Chinese word sequence "C_{adverb}+C_{adject}" that can be matched to a Chinese simple pattern "adv+adj", C_{adverb} has not been aligned yet, and C_{adject} has been aligned to Korean word "$K_{stative\_noun}$". Assume that the context of the Korean word "$K_{stative\_noun}$" in the given sentence is "$K_{general\_adverb}$+$K_{stative\_noun}$+$K_{auxiliary\_verb}$", and only $K_{stative\_noun}$ has been aligned (as in figure 2).

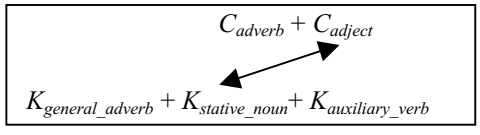

Figure 2. Simple-pattern-based alignment

Calculate the alignment probability of "C_{adverb} → $K_{general\_adverb}$" and " C_{adverb} →$K_{auxiliary\_verb}$" with equation (4).

$$P_{Alignment} = d \times (P_{CPattern} \times P_{CPOS \to KPOS}) \quad (4)$$

where

$P_{Cpattern}$ is the occurrence probability of Chinese pattern "adv+adj" in Chinese tree bank.
$P_{CPOS \to KPOS}$ is the Chinese-Korean POS transferring probability.

[1] In fact, we found that the statistic-based alignment spents the most of the time in word aligning.

[2] *ClassSim(C,K)*: A concept similarity of Chinese class C and Korean class K (Huang & Choi, 1999)

d is a constant that is in inverse proportion to the interval between the Chinese/Korean word to be aligned and the Chinese/Korean reference word that has been aligned.

In the given example, the probability of Chinese adverb transferring to Korean general adverb is 37.9%, when the probability of Chinese adverb transferring to Korean auxiliary verb is 26.0%. All probabilities that bigger than the given threshold will remain until filterring and selecting the best stage.

3.3.6 Filtering and Best Selection

In this stage, we will filter the alignment candidates that we've gotten, and remain only the best one. If the similarities among candidates are very close, then they will all remain as a result of 1:n (one to many) or n:n (many to many) alignment.

For example, the similarity of "[C]下雨(*xiayu*) ⇔ [K]비(*bi*)" and "[C]下雨(*xiayu*) ⇔ [K]내리(*naeri*)" are very close to each other, then it will be considered that "[C] 下雨(*xiayu*)" is aligned to Korean phrase "[K]비 내리 (*bi nari*)".

4 Experiment

We used bilingual dictionary that contains 140,000 items in dictionary-based alignment. And we used 60,000 sentence pairs as training corpus in the statistic-based alignment. In the simple pattern-based alignment, about 300 of simple patterns are employed. There are 120 bilingual sentences in the test set, with 13.7 Chinese words and 9.5 Korean content words (17.9 Korean words) in a sentence pair on average. The sentences of the test set are not contained in the training corpus.

The first experiment is designed to demonstrate the effectiveness of every algorithm independently. The experimental result (Table 2) shows that the dictionary-based alignment shows high precision and low coverage. When we try to rise up the coverage, the precision falls down remarkably. Statistic-based algorithm shows limitation in both of coverage and precision, this result is not surprising because of the reason that we have discussed in our paper. The class-based algorithm is inefficient even than statistic-based algorithm, actually it is out of our imagination, and this result is much more different from the Ker's (1997). The main reason of it is that when we use Dic-coefficient eqution to calculate the Chinese-Korean word similarity, the noise that we have mentioned in our paper is more serious than that is in the English-Chinese. Simple-pattern-based algorithm is not listed below, because this stage aligns words by using the aligned word information, it means that this stage can not work independently. Filtering stage are employed after every algorithm in the below.

Algorithm and Threshold	Coverage	Precision
Dictionary-based, (>0.90)	25.5%	94.4%
(>0.81)	33.7%	89.2%
(>0.66)	45.6%	86.1%
(>0.49)	55.8%	78.4%
Statistic-based, (>0.20)	26.7%	84.6%
(>0.05)	38.1%	72.8%
Class-based, (>1.1)	20.0%	65.7%
(>0.75)	31.6%	57.8%
(>0.5)	47.3%	47.2%

Table 2. Effectiveness of every independent algorithm

In table 2, the recall of the dictionary-based alignment looks quite low, considering the fact that 60% of Korean colloquial words are derived from Chinese words, and most of them share formation similarity. Besides the reasons that we have discussed in session 2.2.1, the experiment shows us that the percentage of the Chinese Korean words in corpus is much lower than the percentage in the dictionary. For example, there are a Korean word "부수다 (*busuda*)" and Chinese Korean word "분쇄하다 (*bunswaehada*) ⇔ [C]粉碎 (*fensui*)" in Korean that corresponds to the same meaning of "break into pieces", but the Korean word "부수다

(*busuda*)" is used more frequently in corpus than the Chinese Korean word "분쇄하다 (*bunswaehada*)".

The second experiment is designed to demonstrate the effectiveness when more than two algorithms are employed together. The experimental result (Table 3) shows that statistic-based algorithm and class-based algorithm are helpful to improve the coverage without the falling down of the precision, when they are used with dictionary-based approach. Though the improvement of the coverage maybe seems not so conspicuous, considering our another experimental result that the upturn of the coverage is only 3% under the same precision when we upgrade our bilingual dictionary from 6,000 items to 140,000, the result in the Table 3 is quite remarkable.

From the experiment result, we can see that to get high precision with higher coverage, using different knowledge is useful - as we have previously stated, using only one or two kinds of information will cause the low correctness inescapably.

The last line of the table 3 demonstrates the effectiveness when all algorithms employed together. Simple-pattern-based alignment is done additionally to its previous stage. Our simple-patterns are mainly gotten by statistical information without enough manual editing. We believe the precision and coverage of this stage will be improved if we check the patterns throughly.

Algorithm and Condition	Coverage	Precision
Dictionary & Statistic-based (dic>0.66 && sta>0.05) || (dic>0.9) || (sta >0.20)	36.5%	92.6%
Dictionary & Class-based (dic>0.66 && cls>0.50) || (dic>0.9)	29.1%	92.7%
Dictionary & Statistic & Class-based (dic>0.66 && sta>0.05) || (dic>0.66 && cls>0.50) || (dic>0.81 && cls>0.35) || (dic>0.9) || (sta >0.20)	39.1%	93.0%
+ Simple-pattern-based alignment	55.1%	89.2%

Table 3. Effectiveness when different algorithms used together

As a result of the experiment, we can see that 60% of Chinese-Korean corresponding relations are 1:1 relations. When including the 1:2 and 2:1 relations, the percentage can raise up to 95%. And in the most of the 1:2 relations, the Chinese words have specific POS and internal structure.

5 Conclusion

This paper clarifies the definition of alignment from the viewpoint of linguistic similarity. Based on our clarified definition, we can easily see that the alignment problem is essentially the problem of bilingual word similarity. We propose that linguistic knowledge would be more efficient than traditional statistical information in the word and phrase alignment, especially between the languages that are not from the same language family. The result of the experiments sustains our proposal to some degree.

We make a linguistic comparison between Chinese and Korean from the viewpoints of character, lexicon, and syntax The lexical and syntactic similarities proposed in the paper exist in the other language pairs also. And the use of such similarities is helpful to raise coverage and precision, especially in the alignment between the languages that do not belong to the same language family.

The coverage is not very high even though we employed serious approachs in our system, we will try to do more works to enhance it. Another work has to be done for the alignment features, they are selected heuristically more than theoretically now. We would like to extend word-level alignment to phrase-level in next step, it will be helpful to the construction of translation patterns.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the "Multilingual Informaton Retrieval" project at the Advanced Information Technology Research Center (AITrc).

This work was also supported by the Korea Science and Engineering Foundation (KOSEF) through the project of "Translation Knowledge Acquisition Through Chinese-Korean Alignment".

References

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra and R.L. Mercer (1993) *The Mathematics of Statistical Machine Translation: Parameter Estimation.* in Computational Linguistics, 19(2), pp. 263-311.

Choi, Feng Chen (1989), *The Lexicon Comparison between Korean and Chinese*, The Press of Yanbian University. (In Chinese)

Church, K. (1993) *Char_align: A Program for Aligning Parallel Texts at the Character Level*. In Proceedings of the 31st Annual Conference of the Association for Computional Linguistics. pp.1-8

Dagan, Ido, Alon Itai, and Ulrike Schwall (1991) *Two languages are more informative than one*. In Processing of the 29th Annual Meeting of the Association for Computational Linguistics, pp.130-173.

Dagan Ido, K.W. Church and W.A.Gale (1994) *Robust Bilingual Word Alignment for Machine-Aided Translation*, In Proceedings of 4th conference on Applied Natural Language Processing (ANLP-94), pp.34-40.

Dan, I. Melaned (1997) *A Portable Algorithm for Mapping bitext Correspondence,* In Proceedings of 35th Conference of the Association for Computational Linguistics

Dice,L.R.(1945) *Measures of the amount of ecologic association between species*. Journal of Ecology, 26:297-302

Gale, W.A. and K.W. Church (1991) *A program for aligning sentences in bilingual corpora*. In Proceedings of the 29th Annual Conference of the Association for Computational Linguistics, pp.177-184.

Frank Smadja, Kathleen McKeown, and Vasileios Hatzsivassiloglou (1996) *Translating collocations for bilingual lexicons: A statistical approach*. In Computational Linguistics, 21(4): pp.1-38.

Fung, Pascale (1995) *A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora*, In Proceeding of the 33th Annual Conference of the Association for Computational Linguistics, pp. 236-243

Fung, Pascale and Lo Yuen Yee (1998) *An IR Approach for Translating New Words from Nonparallel, Comparable Texts*, In Proceeding of the COLING-ACL'98, pp.414-420

Huang, Jin-Xia and Key-Sun Choi (1999) *Automatic Construction of Lexical Classification Net for Two Languages*, In Proceeding of the 11th Conference of the Korean Language and Information Processing, pp.389 – 396 (In Korean)

Ker, Sue J., Jason S. Chang (1997) *A Class-based Approach to Word Alignement*. Computational Linguistics (1997 Volume 23, Number 2 , pp.313 - 343

Kevin McTait, Arturo Trujillo (1999) *A Language-Neutral Sparse-Data Algorithm for Extracting Translation Patterns*, In Proceedings of 8th International Conference on Theoretical and Methodological Issues in Machine Translation, pp. 98 - 108.

Li, Charles N. & Sandra A. Thompson (1981) (Translated by Bak Jeong-Gu,Bak Jong-Han, Baek Eun-Yyi, O Mun-Yi, Coe Yheong-Ha). *Standard Chinese Grammar*. pp. 34- 45

Li, Jun-Jie, Key-Sun Choi (1997) *Design and Implementation of a Chinese-Korean Machine Translation System*. In Proceedings of the 17th International Conference on Computer Processing of Oriental Languages (ICCPOL'97), pp. 400-403,

Martin Kay and Martin Roscheisen (1993) *Text-Translation alignment*. In computational Linguistics, 19(1): pp.121-142.

Mei, Jia-Ju, Yi-Ming Zhu, Yun-Qi Gao, Hong-Xiang Yin, (1983), Tongyici Cilin (Chinese Synonym Forest), ShangHai Press of Lexicon and Books (in Chinese)

Shin, Jung H., Young S.Han and Key-Sun Choi (1996) *Bilingual Knowledge Acquisition from Korean-English Parallel Corpus Using Alignment Method (Korean-English Alignment at Word and Phrase Level)*, In Proceedings of the 15th International Conference on Computational Linguistics, pp.230-235.

Tanaka Takaaki and Yoshihiro Matsuo (1999) *Extraction of Translation Equivalents from Non-Parallel Corpora*, In 8th Proceedings of International Conference on Theoretical and Methodological Issues in Machine Translation, pages 88 - 97.

Turcato Davide (1998) *Automatically creating bilingual lexicons for machine translation from bilingual text*. In Proceedings of the 16th International Conference on Computational Linguistics. pp.1299-1306

Wu, Dekai (1994) *Aligning a parallel English-chinese corpus statistically with lexical criteria*. In Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics (ACL'94), pp.80-87.