USING BILINGUAL SEMANTIC INFORMATION IN CHINESE-KOREAN WORD ALIGNMENT

Jin-Xia Huang[*,**], *Key-Sun Choi*[*]

[*]KORTERM, Computer Science Division
Korean Advanced Institute of Science and Technology
373-1 Yusong-ku Taejon 305-701 Korea

[**]Yanbian University of Science and Technique, Beishan Street, Yanji, 133001, P.R.China

Email: {hgh,kschoi}@world.kaist.ac.kr

ABSTRACT

This pape clarifies the definition of alignment from the viewpoint of linguistic similarity. We propose new method for the alignment betwee the languages that do not belong to the same language family. On the contrary to most of the previously proposed methods that rely heavily on statistics, our method attempts to use linguistic knowledge to overcome the problems of statistical model.

1. INTRODUCTION

1.1 Previous Works

Bilingual corpus provides more information than monolingual one (Dagan, 1991). In recent years many works have been done on word alignment after the research on section, paragraph, sentence, and phrase level alignment. Word alignment works not only for th automatic construction of bilingual dictionaries and other useful resourc s, but also for the various applications such as machine translation (MT) and word sense disambiguation.

Statistical approach has been used as main technique in most alignment systems (Gale, 1991; Brown, 1993; Dagan, Church, Gale, 1993). Correlation information of bilingual lexicon is mainly employed in statistical approach, and other information, such as character information or position information is used additionally (Brown, 1993; Dagan, 1993). In the alignment of some language pairs that do not belong to the same language family, word list (Wu, 1994) or functional word information (Shin and Choi, 1996) was employed additionally.

In contrast to the systems that mainly rely on statistical approach, the English-Chinese alignment system developed by Ker (1997) used a class-based algorithm in its word alignment, and the syste did not use any statistical technique. The result show ed that this new approach could overcome the lower coverage of the statistical approach whil gaining high precision, even if his system was only based on small test set. Ker's work clews us the feasibility of pure linguistic approach in the resol utio of the alignment problem.

Phrase level alignment is one of the most difficult problems in word alignment. Some statistical approaches have been used to resolve the problem, but most of them get hold of not linguistic phrases but statistical ones (Wu, 1994; Shin & Choi, 1996). Using parser maybe a solution to this problem (Turcato, 1998), but it can cause other problems such as data sparseness because of the syntax diversity between th two languages. For the improvement of correction, some research restrained the alignment object to som special phrase such as compound noun (Kupiec, 1993; Kumano, 1994).

[1] The research described in this paper is sponsored by Advanced Information Technology Research Center (AITrc) and Korean Science and Engineering Foundation (KOSEF)

Phrase level alignment is especially much more difficult when the language pairs do not belong to th same language family, becaus the syntax structure and the lexical formation are much more different in this case there are more cases out of many-to-many one to one correspondenc e relation between these languages. Because the information employed in conventional alignment system is not enough in such languages, it causes the decline in both of coverage and precision (Shin & Choi, 1996).

1.2 Our Proposal

In most of previous research es, the definition of alignment was represented as "align word (text, phrase section etc) to its translation" (Gale, 1991; Shin & Choi, 1996). It seems like the concept of alignment is so obvious that no one have concer ed 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 a work on finding the translation of the source language. "Word alignment" should find out the object that shares the highest semantic similarity with the source word. When there is more than one candidate, system should find the object that shares the highest syntactic similarity among th candidates. If word level alignment is impossible because of the linguistic peculiarity, alignment object can be expand ed to phrase, which contains the least words and shares maximum semantic similarity with source word or phrase.

From the above concrete definition, we can easily find that alignment problem is essentially the problem of bilingual word similarity calculation.

The traditional statistical approach has been testified to be effective in the resolution of alignment problem, it shows that statistical information reflects word similarity in some stage. But this approach get good result mainly betwee the languages that belong to the same language family (Brown, 1993; Dagan, 1993) and shows limitation in coverage even after training with extremely large bilingual corpus (Ker, 1997). To the languages that do not belong to the sam language family, the statistical approaches have shown limited coverage and low correctness, even after the employment of additional information (Shin and Choi, 1996; Turcato, 1998). This result is not surprising becaus statistical approach is just an indirect way to get the word similarity, and the information employed in the previous syste s, such as statistical information, word position, functional word or rule, is indirect and vague knowledge in reflecting word similarity.

Then, what is the more direct information in getting bilingual word similarity? In monolingual processing, some resources such a s dictionary, thesaurus and WordNet have been customar ily used. Alignment needs bilingual information, so we attempt to use bilingual dictionary instead of monolingual dictionary . Thesaurus and WordNet have almost never been used in bilingual alignment excepts 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.

Besides this, we believe that there are character similarity, lexicon similarity, syntactic similarity between some languages (e.g., Chinese and Korean). At least, above linguistic knowledge is more close to bilingual word similarity than something lik statistical information or word position information or so. We will discuss about it in next section.

2. LINGUISTIC COMPARISON BETWEEN CHINESE AND KOREAN

We will compare the linguistic property of Chinese and Korean from the three viewpoints - character, lexicon, and syntax. At the end of the comparison, we will discuss bilingual word similarity and relativ information.

2.1 Character Comparison

In theory, there is always at least one Korean character corresponding to any Chinese characters. Practically, there are some rarely used Chinese characters have no corresponding Korean characters. Most of the Chinese characters have exactly one corresponding Korean character (Ex, "[C]'名[*Ming*]' ⇨ [K]'名[*myeong*]'", "[C]'快[*Kuai*]' ⇨ [K]'快[*kwae*]'"), and seldom of them have 2 or more ones (Ex: "[C]'便[*Bian*]' ⇨ [K]'변[*byeo*]', '편[[*pyeon*]'").

We constructed a Chinese-Korean Character Transfer Table (CKCT Table) to reflect the correspondence relatio . There are totally 436 different Korean characters corresponding to 6763 Chinese characters that are listed in GB2313-80 Chinese code table.

2.2 Lexicon Comparison

In history, Korean language was influenced by Chinese languag especially in the formativ process of lexicon. We will look at th lexical similarity betwee the two languages from three viewpoints - lexical formatio , part of speech (POS) and lexical intra-structure.

(1) Lexical formation

60% of Korean words are derived from Chinese words. If these Korean lexicons are transformed into corresponding Chinese characters or vice verse, they will have similar form and meaning (Ex, "[C] 和平[*Heping*]⇔[K][*pyeonghwa*](平和[*Pinghe*])(⇔[E]peace)","[C]办公室[*Bangongshi*]⇔[K][*samusil*](事务室[*Shiwushi*]) (⇔[E]office)". (Li and Choi, 1997)

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

(2) POS

POS similarity indicates the regularity betwee 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 translati of the word is also a pronoun.

POS similarity has been gotten attention by computational linguistic researchers long before and has been made use in several previous alignment systems (Dagan, Church, Gale, 1994; Shin and Choi, 1996).

(3) Lexical intra-structure

Different from other language words, Chinese words have intra-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)", we'd like to say that intra-structure of word "下雨[*Xiayu*](rain)" is "verb+noun".

We found that in most cases of one-to-many (1:n) correspondence, Chinese words have some specific POS and intra-structure. And the Korean phrases will hold similar syntactic structur s whil the source Chinese words have the same intra-structure. We name it "lexical intra-structure similarity" in our paper.

For example, in next two transfer examples, the Korean phrases have the same syntactic structure while the corresponding Chinese words have the same intra-structure: "[C][v➔v+n]⇔[K][Vp➔n+aux v+termination]"

fall rain (⇨ rain)
"[C]下 + 雨 [*Xiayu*]"
"[K]비+가 오+다[*biga oda*]":
rain come (⇨ rain)

return home (⇨ go home)
"[C]回 + 家[*Huijia*]"
"[K]집+에 가+다[*jibe gada*]" :
home go (⇨ go home)

In above example, "[C][v➔v+n]" is Chinese word intra-structure, "[K][Vp➔n+aux v+termination]" is syntactic structure of Korean phrase "[C]下雨[*Xiayu*]([E]rain)" and "[C] 回家[*Huijia*] ([E]go home)" ar Chinese words, "[K][*biga oda*] ([E]rain)" and "[*jibe gada*] ([E]go home)" ar corresponding Korean phrase.

The lexical structure transfer rules (*e.g.*, "[C]verb➔verb+noun⇔[K]Vp➔noun+aux verb +termination") can be constructed semi-automatically . Chinese word intra-structure can be partly gotten from the Grammatical Knowledge-base of Contemporary Chinese (Yu, 1998) and Chinese dictionary.

(4) Lexical sense

Even if two languages do not belong to the same language family, their lexicon has semantic similarity because the objects they want to describe are the same world. One of the best examples about semantic similarity between two languages is bilingual dictionary, almost all of the source word s have their translation in target language.

Concept similarity is closely related to semantic similarity. Formatio , POS and intra-structure of lexicon all r flect semantic similarity as we have discussed above. Besides them, the syntactic similarity (that we will discuss below) and statistical information that has been used in most alignment system all reflect semantic similarity to some extent.

2.3 Syntactic Comparison

Word position (*e.g.,*between English and French), functional word (*e.g.,* between Korean and English) and POS information all reflect syntactic information, and they have been used with statistical approach in previous works (Gale, 1991; Brown 1993; Shin & Choi, 1996). But in the alignment of Chinese and Korean, word position in sentence is not correct enough because their word order s are quite different. Chinese is SVO type languag whil Korean is SOV one, and both of their word orders ar quite flexible. Additionally, Chinese word order is reflected by semantic element more than by syntactic on (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 patter s that contain no embedded structure (right hand side of next examples).

new book

"[C]新/adj 书/noun[*Xin Shu*]" ([C]adj noun)

"[K]새/adj 冊/noun[sae caeg]" ([K]adj noun)

new book

this person

"[C]这/pron 人/noun[*Zhe Ren*]" ([C]pron noun)

"[K]이/pronoun 사람/noun[*i saram*]" ([K]pron noun)

this person

discuss discuss(⇨Let' have a debate) ([C]v1 v1)

"[C] 讨论/v 讨论/v [*Taolun Taolun*]"

"[K] 議論/noun+하여 보자 [*yinonha'yeo boja*] "

discussion do let's (⇨Let's have a debate)

([K]noun+하여 보자)

Using simple transfer pattern will be more correct than only using word position in sentence or POS information. And it will be benefit to over come the data sparseness problem than using parser or sentenc transfer pattern.

3.CHINESE-KOREAN WORD ALIGNMENT

3.1 Alignment Object

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

3.2 Resource an Information used in the System

Figure 1 shows the linguistic resources we used in our system and information they can provide.

Resource	Information
CKCT Table	Similarity of lexical formation
Bilingual Dictionar	Semantic similarit
Bilingual Class-Net	Conceptual Similarit
Grammatical Dictionary	Similarity of lexical intra-structure
Simple Transfer Pattern	Syntactic Similarit
Bilingual Corpus	Correlation information of bilingual words

Figure 1. Resource and information used in Chinese-Korean alignment system

Bilingual dictionary provides us the target words that have highest semantic similarity with the source words. Bilingual Class-Net is constructed with Korean and Chinese m onolingual thesaurus , and it provides us the conceptual similarity between Chinese and Korean words (Huang & Choi, 1999). CKCT Table helps us to get similarity of lexical formatio . And grammatical dictionary provides us some grammatical information of source lexicons and it is helpful in getting the lexical structure similarity Simple transfer pattern makes use of simple pattern information in getting syntactic similarity. Finally, bilingual corpus provides us correlation information of bilingual words as is quite well known.

3.3 Alignment Algorithm

We use Chinese and Korean POS tagger as preprocessor of our system. The Chinese-Korean alignment system architecture is shown in Figure2.

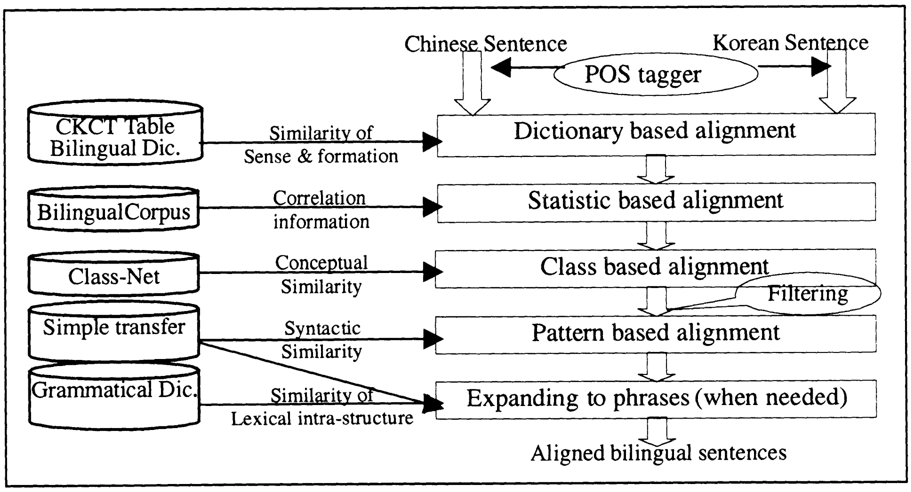


Figure2. Chinese-Korean Word Alignment System

Figure3 is an example that will be used for our alignment algorithm.

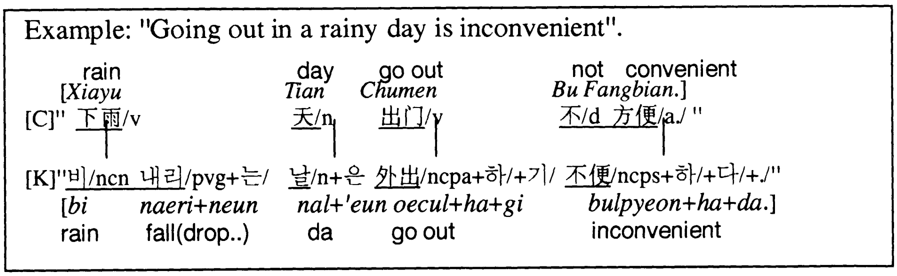

Example: "Going out in a rainy day is inconvenient".
rain
[Xiayu
[C]" 下雨/v
day
Tian
天/n
go out
Chumen
出门/v
not convenient
Bu Fangbian.]
不/d 方便/a./ "
[K]"비/ncn 내리/pvg+는/ 날/n+은 外出/ncpa+하/+기/ 不便/ncps+하/+다/+./"
[bi naeri+neun nal+'eun oecul+ha+gi bulpyeon+ha+da.]
rain fall(drop..) da go out inconvenient

Figure 3. Chinese-Korean alignment example

Stage1: 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 formatio . Using CKCT table, and transfer the given Chinese word to its corresponding Korean word by transferring characters one by one, and then add it to empty set kci.

1. Search the Korean translation words of the Chinese word c_i from bilingual dictionary. Add them to the set k_{ci}. We consider one Korean phrase that contains no space as one word, and delete postfix of Korean word if the word is verb, adverb or adjective.
2. Calculate similarity of c_i and k_j - *WordSim(*c_i*,*k_j*)* with equation (1). If *WorSim(*c_i*,*k_j*)*$>t_1$(t_1 is threshold), then save the result to set *Similarit* $_{ij}$.

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

where

c = Chinese word of given sentence
k_c = Korean lexical set correspondent to C though CKCT and bilingual dictionary.
k_{Ci} = element i of set k_c
k = Korean morpheme of given sentence
$|\,k\,|$ = totel number of the characters in k
$|\,k_{Ci}\,|$ = totel number of the characters in ith element of set k_C
d = d<1.00, if k_{Ci} and k have diffirent POS tagger, or $|c|$=1 and $|k|$=1,
d = 1.00, otherwise

For the example of figure 3, two word pairs ("出门[*Chumen*] (go out)" and "外出[*oecul*] (go out)", "天[*Tian*] (day)" and "날[*nal*] (day)") will be aligned correctly in stage 1.

1. From CKCT Table:
   "出门[*Chumen*] (go out)"="出門[*culmun*][2]"
   Set k_c = {"出門"}
2. From Bilingual dictionary,
   "出门[*Chumen*] (go out)" = "집을 떠나다[*jib'eul ddeonada*](leave home),외출하다[*oeculhada*](go out), 밖에 나가다[*bagg'e nagada*](go out), 出嫁하다[*culgahada*](leave home), 시집가다[*sijibgada*](take a husband), 結婚하다[*gyeolhonhada*](marry)"

   Set $k_{c'}$ = {"出門[*chulmun*]", "집을[*jib'eul*]", "떠나[*ddeona*]", "外出[*oecul*]", "밖에[*bagg'e*]", "나가[*naga*]", "出家[*culga*]", "시집가[*sijibga*]", "結婚[*gyeolhaon*]"}[3]
3. Calculate WordSim("出门[*ChuMen*](go out)", "外出[*oecul*](go out)") as follows, and add the result to set *WordSim*$_{ij}$:
   Set *WordSim*$_{ij}$ = {1.0}

WordSim(出门[*Chumen*] (go out), 外出[*oecul*] (go out))[4]

$$= 1.0 \times \max\left\{\frac{2\times|\text{출문}\cap\text{외출}|}{|\text{출문}|+|\text{외출}|}, \frac{2\times|\text{집을}\cap\text{외출}|}{|\text{집을}|+|\text{외출}|}, \frac{2\times|\text{떠나}\cap\text{외출}|}{|\text{떠나}|+|\text{외출}|}, \frac{2\times|\text{외출}\cap\text{외출}|}{|\text{외출}|+|\text{외출}|}, \frac{2\times|\text{밖에}\cap\text{외출}|}{|\text{밖에}|+|\text{외출}|}, \frac{2\times|\text{나가}\cap\text{외출}|}{|\text{나가}|+|\text{외출}|}, \frac{2\times|\text{출가}\cap\text{외출}|}{|\text{출가}|+|\text{외출}|}, \frac{2\times|\text{시집가}\cap\text{외출}|}{|\text{시집가}|+|\text{외출}|}, \frac{2\times|\text{결혼}\cap\text{외출}|}{|\text{결혼}|+|\text{외출}|}\right\}$$

$$=1.0\times\max\left\{\frac{2\times1}{2+2},\frac{2\times0}{2+2},\frac{2\times0}{2+2},\frac{2\times2}{2+2},\frac{2\times0}{2+2},\frac{2\times0}{2+2},\frac{2\times1}{2+2},\frac{2\times0}{3+2},\frac{2\times0}{2+2}\right\}$$

$$=1.0\times\max\{\ 0.5,0,0,1,0,0,0.5,0,0\}$$

$$= 1.0\times1$$

[2] "出門[*culmun*]" is a meaningless character string.

[3] In the set, some phrases are words ("外出[*oecul*]" - go out), some are phrases that part of Korean sub sentence("집을[*jib'eul*]" - part of "집을 떠나[*jib'eul ddeona*](leave home)"), and some are meaningless character string ("出門[*culmun*)").

[4] About the phrase pronunciation and the meaning in English, please refer to the item 2 just above and the footnote 3.

Stage2: Statistic based alignment

We will align words that have high co-occurrence, but share lo formation similarity or can not been found in bilingual dictionary. We employed t-score to reflect correlation of bilingual words as equation (2), and if the result bigger than threshold, then save it to set $Similarity_{ij}$.

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

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

To example of figure 3, "下雨 [Xiayu]" and "비[bi]" will be aligned correctly (but not completely) in stage 2.

Stage3: Class-based alignment

Chinese Tongyici Cilin (Synonym Forest, CILIN), Korean thesaurus and Class-Net will be employed in stage 3. Class-Net is automatically constructed with Chinese *Tongyici Cilin* and Korean thesaurus (Huang & Choi, 1999). Let's look at th examples of ClassNet:

ClassSim([*C*]Ab01, [*K*]12020)[5] = 1.0;
ClassSim([*C*]Ab01,[*K*]12040)[6] = 0.9;
ClassSim([*C*]Ab01,[*K*]12340)[7] = 0.6...
Where, *ClassSim(C,K)* is concept similarity of Chinese class *C* and Korean class *K* .

The algorithm of class-based alignment is as follows:

1. Search the class C_i of given Chinese word c_i in Cilin, get all of the Chinese words that contained in the class C, transfer them to corresponding Korean word s by transferring characters one by one, and then add them to empty set K_{Ci}.

2. Calculate similarity of c_i and given Korean word k_j with equation (1) as in stage 1. If $WordSim(c_i,k_j)> t_1$(t_1 is threshold), then save the result to set $Similarit_{ij}$, go to next word c_{i+1}. Else go to step 3.

3. Get concept similarity from Class-Net. If $ClassSim(C_i,K_j)> t_2$(t_2 is threshold), then save the result to set $Similarit_{ij}$.

To example of figure 3, "方便[Fangbian](convenient)" and "不便[bulpyeon] (inconvenient)" will be aligned correctly (but not completely) in stage 3.

1. From *Chinese Cilin*, search the class of "方便[FangBian](convenient)", getting all of the Chinese words contained in the class:
   Set C_i = {Ed48, Ed49, Ef44, Je10}
   = {"便利 [*Bianli*](convenient)", "便当[*Biandang*](convenient)",
   "不便[*Bubian*](inconvenient)", ...}

2.Transfer getting Chinese words to corresponding Korean words:
   Set K_{Ci} = {"便利[*pyeonri*](convenient)", "便當[*pyeondang*](meaningless)",
   "不便[*bulpyeon*](inconvenient)", ...}

3. Calculate similarity of "方便[*fangbian*](convenient)" and "不便[*bulpyeon*] (inconvenient)" with equation(1) given in first stage.
   WordSim(方便, 不便)[8]
   =Max(WordSim(便利,不便),WordSim(便當,不便), WordSim(不便,不便), ...)
   = MAX(0.5, 0.5, 1.0, ...) = 1.0

The result is bigger then threshold, save it to $Similarit_{ij}$.

[5] Ab01: man, woman; 12020: human
[6] Ab01: man, woman; 12040: man and woman
[7] Ab01: man, woman; 12340: person, characte
[8] About the word pronunciation and the meaning in English, please refer to the item 2 just above.

- Filtering before Stage 4:
To raise the precision, system will filter the alignment result before stage 4 with some heuristics.

For example, if there is one candidate to one Chinese word, and the similarity is high enough (bigger than threshold ts), then it will be marked "1" (level 1, means correctly aligned). And if the similarity is high but not high enough (bigger than threshold t1 or t2 but smaller than threshold ts), it will be marked "2" (level 2, means not finished yet).

If there are more than two candidates to one Chinese word, the best one will be marked selected, and if the similarities are very close to each other, then them will be remained all.

Stage4: Pattern based alignment
We will align the words that have not bee aligned yet until stage 4 using simple transfer pattern. In this stage, aligned word information and word position information will be employed, and only the word position that inside the range of simple pattern will b considered useful.

Let's look at the algorithm of stage4 with an example:

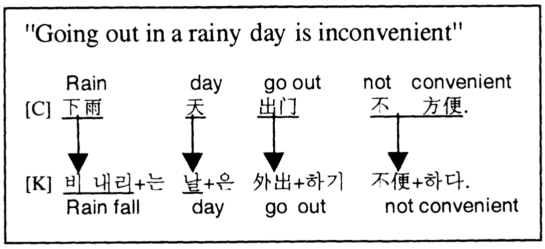
To example of figure 3, "[C]不+方便 [*Bu+Fangbian*](not+convenient)" and "[K]不便[*bulpyeon*](inconvenient)" will be aligned completely in stage 4.

1. Get the first unaligned content word "[C]不 [Bu](not)" from Chinese sentence.
2. Match the phrase '[C]不/adv 方便/adj[*Bu Fangbian*](not convenient)' to Chinese simple pattern "[C]adv+adj". We can found that " [C]方便 [*Fangbian*](convenient)"has been aligned already, so go to next step. (If there is no aligned word inside the pattern, go to step 1.)
3. Get th corresponding word of "[C]方便 [*Fangbian*](convenient)" in Korean sentence: "[K]不便[*bulpyeon*] (inconvenient)". We can found that the phrase pairs " [C]不/adv 方便/adj[*Bu Fangbian*](not convenient) ⇔ [K]不便/stative noun [*bulpyeon*](inconvenient)" can be matched to transfer pattern "[C]adv+adj ⇔ [K]stative noun". So there is at least possibility that they may b aligned.

Stage5: Expanding to phrases when needed
We will try to do 1:n, n:1 and n:n alignment in this stag with some heuristics.

1. To some Chinese words that have both of specific POS and intra-structure, check if there is possibility of 1:n corresponding relation. (Ex, "[C]下雨[*Xiayu*](rain)" ⇔ "비 내리[*bi naeri*](rai fall)")
2. To the Chinese and Korean words that have above two alignment candidates, check if there are possibility of n:1 or 1:n corresponding relation, simple transfer pattern will be employed in this step.
3. To the possible relations of 1:n, n:1 and n:n, if the similarities of candidates are very closed to each other and they are all high enough, it will be considered as correct alignment. Otherwise, the relation will be deleted.

As the result of stage 5, the sentence pairs "下雨天出门不方便.[*Xiayu Tian Chumen Bu Fangbian*]" ⇔ "비 내리는 날은 外出하기 便하다.[*bi naerineun nal'eun oeculhagi bulpyeonhada*]" (⇔ "Going out in a rainy day is inconvenient") will be aligned as follows:



4. EXPERIMENT

Our Chinese-Korean alignment system is still under construction.

We used bilingual dictionary that contains 6,000 items in dictionary-based alignment. And used 60,000 sentence pairs as training corpus in statistic based alignment. In Pattern based alignment, for about 300 pairs of simple transfer patterns are employed.

Figure 4 is one of experimental results:

Stage	Recall	Precision
Dictionary based alignment	22.2%	95.1%
Statistic based alignment	30.3%	91.2%
Class based alignment	-	-
Pattern based alignment	47.6%	90.1%
Expanding to phrase	-	-

Figure 4. Experiment result

The recall of dictionary based alignment looks quite low while consider the fact that 60% of Korean words are derived from Chinese words (Chinese Korean words), and most of them share formatio similarities. It because that, the percentage of the Chinese Korean words in corpus is much more lower then the percentage in dictionary For example, there are Korean word "부수다[*busuda*](break)" and Chinese Korean word "粉碎하다[*bunswaehada*][粉碎](break)" in Korean that corresponding to the same meaning of "break into piece ", but the Korean word " 부수다[*busuda*](break)" are more frequently used in corpus then the Chinese Korean word "분쇄하다[*bunswaehada*][粉碎](break)".

As a result of experiment, we could see that 60% of Chinese Korean corresponding relations are 1:1 relatio . If include the 1:2 and 2:1 relation s, the percentage can rise to 95%. And in most of the 1:2 relations, the Chinese words hav specific POS and intra-structure.

5. CONCLUSION

This pape clarifies the definition of alignment from the viewpoint of linguistic similarity. We can easily see that the alignment problem essentially the problem of bilingual word similarity calculation fro the clarified definition. Based on the definition, we proposed that linguistic knowledge would be stronger than traditional statistical approach in word and phrase alignment, especially between the languages that do not belong to the same language family. The experiment result of our alignment system partly sustains our proposal.

We make a linguistic comparison between Chinese and Korean from the viewpoints of character, lexicon, and syntax. The lexical and syntactic similarities that proposed in the paper exist in other language pairs too. 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.

Acknowledgments: We would like to thank Mr. Byeongju Kang, Miss Hyijeong Song, Miss Yeongmi Song and Miss Jihyeon Kim, for their fruitful collaboration and help.

6. REFERENCES

[1] 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, pages 130-173.

[2] Gale, W.A. and K.W. Church, 1991, "A program for aligning sentences in bilinbual corpora". In Proceedings of the 29th Annual conference of the Association for Computational Linguistics, pp.177-184.

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

[4] Dagan, Ido, Kenneth W. Church, and William A.Gale.1993. "Robust bilingual word alignment for machine aided translation". In Proceedings of the 29th the Workshop on Very Large Corpora: Academic and Industrial Perspectives, pp.1-8.

[5] 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, pages 80-87.

[6] Shin, Jung H., Young S.Han and Key-Sun Choi.1996. "Bilingual Knowledge Acquisition fro 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, pages 230-235.

[7] Ker, Sue J., Jason S. Chang. 1997. "A Class-based Approach to Word Alignement. Computational Linguistics" 1997 Volume 23, Number 2 , P313 - 343

[8] Turcato, Davide 1998. "Automatically creating bilingual lexicons for machine translation fro bilingual tex ". In Proceedings of the 16th International Conference on Computational Linguistics. pp. 1299-1306

[9] Kupiec, Julian.1993. "An algorithm for finding noun phrase correspondences in bilingual corpora". In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp.17-22

[10] Kumano, Akira and Hideki Hirakawa. 1994. "Building an MT dictionary from parallel texts based on linguistic and statistical information ". In Proceedings of the 15th International Conference on Computational Linguistics, pp. 76-81.

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

[12] 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

[13] Dagan Ido, K.W. Church and W.A.Gale. 1994. "Robust Bilingual Word Alignment for Machine-Aided Translation", In Proceedings of Fourth conference on Applied Natural Language Processing (ANLP-94), Stuttgart, Germany, pp.34-40.

[14] Yu, Shiwen, Xuefeng Zhu, Hui Wang, Yunyun Zhang, 1998. "The Grammatical Knowledge-base of contemporary Chinese (Xiadai Hanyu Yufa Xinxi Cidian Xiangjie)". The press of Tsinghua University. (in Chinese)

[15] Li, Charles N. and Sandra A. Thompson. 1981. (Translated by Jeong-Gu Bak, Jong-Han Bak, Eun-Yi Baek, Mun-Yi O, Yheong-Ha Coe in 1989), "Standard Chinese Grammar. pp.34-45

[16] Huang, Jin-Xia, Key-Sun Choi, 1999. "Automatic Construction of Lexical Classification Net for Two Languages", In Korean Language and Information Processing '99 (in Korean)

[17] Dice,L.R.1945. "Measures of the amount of ecologic association between species". Journal of Ecology, 26:297-302

[18] Mei Jiaju, Yiming Zu, Yunqi Gao, Hongxiang Yi , "Chinese Tongyici Cilin (Synonym Forest, CILIN)", 1983,