Bilingual Knowledge Acquisition from Korean-English Parallel Corpus Using Alignment Method (Korean-English Alignment at Word and Phrase Level)

Jung H. Shin and Young S. Han* and Key-Sun Choi
Department of Computer Science
Korean Advanced Institute of Science and Technology
Taejon, 305-701, Korea
*Department of Computer Science
Suwon University
Kyungki, 445-743, Korea
email: jhshin@stissbs.kordic.re.kr

Abstract

This paper suggests a method to align Korean-English parallel corpus. The structural dissimilarity between Korean and Indo-European languages requires more flexible measures to evaluate the alignment candidates between the bilingual units than is used to handle the pairs of Indo-European languages. The flexible measure is intended to capture the dependency between bilingual items that can occur in different units according to different ordering rules. The proposed method to accomplish Korean-English alignment takes phrases as an alignment unit that is a departure from the existing methods taking words as the unit. Phrasal alignment avoids the problem of alignment units and appease the problem of ordering mismatch. The parameters are estimated using the EM algorithm. The proposed alignment algorithm is based on dynamic programming. In the experiments carried out on 253,000 English words and its Korean translations the proposed method achived 68.7% in accuracy at phrase level and 89.2% in accuracy with the bilingual dictionary induced from the alignment. The result of the alignment may lead to richer bilingual data than can be derived from only word-level aligments.

1 Introduction

Studies on parallel corpus consisting of multilingual texts are often guided with the purpose to obtain linguistic resources such as bilingual dictionary, bilingual grammars (Wu 1995) and translation examples. Parallel texts have proved to be useful not only in the development of statistical machine translation (Brown et al. 1993) but also in other applications such as word sense disambiguation (Brown et al. 1991) and bilingual lexicography (Klavans and Tzoukermann 1990). As the parallel corpora become more and more accessible, many researches based on the bilingual corpora are now encouraged that were once considered impractical.

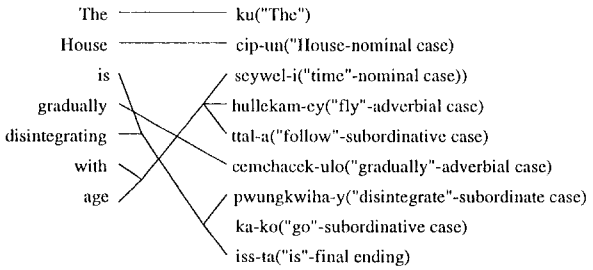

Figure 1: An example of typical Korean-English alignment.

Alignment as a study of parallel corpus refers to the process of establishing the correspondences between matching elements in parallel corpus. Alignment methods tend to approach the problem differently according to the alignment units the methods adopt. Of various alignment options, the alignment of word units is to compute a sequence of the matching pairs of words in a parallel corpus.

Figure 1 show the aligned results of a parallel corpus that was originally paired in a sentence level. In figure 1, the right-hand side of pair-wise alignment is the corresponding Korean words. Described in the parentheses on the right of each Korean word are corresponding English meaning and syntactic functions of the word.

The existing methods for the alignment of Indo-European language pairs such as English and French take words as aligning units and restrict the correspondences between words to be one of the functional mappings (one-to-one, one-to-

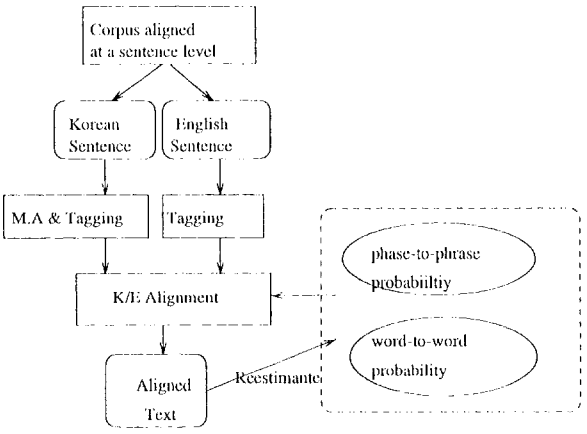

Figure 2: Overview of the proposed alignment method.

many) (Brown et al. 1993, Smadja 1992). These methods made extensive use of the position information of words at matching pairs of sentences, which turned out useful (Brown et al. 1993). The structural similarity in word order and units between English and French must be one of the major factor to the success of the methods.

The alignment of the pairs of structurally dissimilar languages such as Korean and English requires different strategy to compensate the lack of structural information such word order and to handle the difference of alignment units.

An early attempt to align Asian and Indo-European language pairs is found from the work by Wu and Xia (1994). Their result is promising with the demonstration of high accuracy of learning bilingual lexicon between English and Chinese for frequently used words without the consideration of word order. The Chinese-English alignment consists of segmentation of an input Chinese sentence and aligning the segmented sentence with the candidate English sentence. The generation of segments to be aligned is an additional problem to the decision of aligning units before the alignment takes place. Wu and Xia (1994) used bilingual dictionary to segment the sentence, but the selection of segment candidates is hard to make with reliable accuracy. The bilingual dictionaries are not always available and take considerable resources to build.

The method we suggest integrates the procedures to solve the two critical problems: deciding aligning units and aligning the candidates of different word orders and accomplishes the alignment without using any dictionary.

The proposed alignment method assumes a preprocessing step before iterative applications of alignment step as is illustrated in figure 2. Part-of-speech tagging is done before the actual alignment so that the word-phrases (a spacing unit in Korean) may be decomposed into proper words and functional morphemes and the Korean and English words may be assigned with appropriate tags.

The proposed alignment is done first for phrase pairs and then word pairs that eventually induces the bilingual dictionary. The alignment method is realized through the reestimation of its probabilistic parameters from the aligned sentences. In particular, the parameters account for the coocurrence probilities of bilingual word pairs and phrase pairs. The repetitive application of the alignment and reestimation leads to a convergent stationary state where the training stops.

In the following section, our proposed method for aligning Korean-English sentences is described and parameter reestimation algorithm is explained. Section 3 summarizes the results of experiments and Conclusion is given in section 4.

2 Korean/English Alignment Model

2.1 English/French alignment model

To define $p(\mathbf{f}|\mathbf{e})$, the probability of the French sentence $\mathbf{f}$ given the English sentence $\mathbf{e}$, Brown et al. (1991) adopted the translation model in which each word in $\mathbf{e}$ acts independently to produce the words in $\mathbf{f}$. When a typical alignment is denoted by $\mathbf{a}$, the probability of $\mathbf{f}$ given $\mathbf{e}$ can be written as the sum over all possible alignments (Brown et al. 1991)

$$p(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} p(\mathbf{f}, \mathbf{a}|\mathbf{e}) \quad (1)$$

Given an alignment $\mathbf{a}$ between $\mathbf{e}$ and $\mathbf{f}$, Brown et al. (1991) has shown that one can estimate $p(\mathbf{f}, \mathbf{a}|\mathbf{e})$ as the product of the following three terms (Berger et al. 1995).

$$p(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{i=1}^{|\mathbf{f}|} p(n(e_{a_i})|e_{a_i}) \prod_{i=1}^{|\mathbf{f}|} p(f_i|e_{a_i}) d(\mathbf{f}, \mathbf{a}|\mathbf{e}) \quad (2)$$

In the above equation, $p(n|e)$ denotes the probability that the English word e generates n French words and $p(f|e)$ denotes the probability that the English word e generates the French word f. $d(\mathbf{f}, \mathbf{a}|\mathbf{e})$ represents the distortion probability that is about how the words are reordered in the French output.

In the above methods, only one English word is related to one or n French words. The distortion probabilities are defined on the positional relations such as absolute or relative positions of matching words.

2.2 Characteristics of Korean/English alignment

Unlike the case of English-French alignment, Korean and English have different word units to

Table 1: The result of manual analysis about matching unit

Korean words	English words	Ratio
1	1	33.8%
2	1	28.1%
3	1	9.7%
1	2	7.3%
etc.	etc.	11.1%

be aligned, for an English sentence consists of words whereas a Korean sentence consists of word-phrases (compound words). Typically a word-phrase is composed of one or more content words and postpositional function words.

A Korean word is usually a smaller unit than an English word and a word-phrase is larger than an English word. For this reason the exact match as in English-French pair is hard to establish for the case of Korean-English (Shin et al. 1995). Consequently word-to-word or word-to-word-phrase alignment between Korean and English will suffer from unit mismatch and low accuracy. The complication of unit mismatch often implies the need of non-functional alignment such as many-to-many mapping. Non-functional mapping may also occur in the English-French case, but with much less frequency.

The table 1 shows the degree of mismatch between English words and Korean words that are analyzed by our automatic POS tagger and morphological analyzer. When we checked randomly selected 200 sentence pairs by hand, only 33.8% of all pairs have one-to-one correspondences between English words and Korean words.

2.3 Korean to English Alignment

In this section, we propose a Korean to English alignment method that aligns in both word and phrase levels at the same time. First, we introduce the method in word-to-word alignment, and then extend it to include phrase-to-phrase alignment.

By definition, a phrase in this paper refers to a linguistic unit of more general structure than it is recognized in general from the terms, noun and adverb phrases. A phrase is any arbitrary sequence of adjacent words in a sentence.

2.3.1 Base Method (using only word-to-word correspondences)

In the development of our method, we follow the basic idea of statistical translation proposed by Brown et al. (1993). To every pair of sentences of $\mathbf{e}$ and $\mathbf{k}$, we assign a value $p(\mathbf{e}|\mathbf{k})$, the probability that a translator will produce $\mathbf{e}$ as its translation of $\mathbf{k}$, where $\mathbf{e}$ is a sequence of English words and $\mathbf{k}$ is a sequence of Korean words.

$$p(\mathbf{e}|\mathbf{k}) = \prod_{j=1}^{n} \sum_{i=0}^{m} p(e_j|k_i) \tag{3}$$

In equation 3, n and m are the number of words in the English sentence $\mathbf{e}$ and its corresponding Korean sentence $\mathbf{k}$ respectively. e_j and k_i are the aligning unit between English sentence $\mathbf{e}$ and Korean sentence $\mathbf{k}$. e_j represents j-th word in English sentence and k_i represents i-th word in Korean sentence. For example, in Figure 1 English word "the" is e_1 and Korean word "ku" is k_1.

2.3.2 Proposed Method (Extended Method)

The base method of word level alignment is extended with phrase-level alignment that overcomes the difference of matching unit and provides more opportunity for the extraction of richer linguistic information such as phrasal-level bilingual dictionary. To cope with the data sparseness problem caused by considering all possible phrases, we represent phrases by the tag sequences of their component words.

If an English sentence $\mathbf{e}$ and its Korean translation $\mathbf{k}$ are partitioned into a sequence of phrases p_e and p_k of all possible sequences $\mathbf{s}(\mathbf{e}, \mathbf{k})$, we can write $p(e|k)$ as in equation 5 where p_e and p_k are phrase sequences and $\mathbf{a}(p_e, p_k)$ denotes all possible alignments between p_e and p_k.

$$\begin{aligned} & p(\mathbf{e}|\mathbf{k}) \\ &= \sum_{<p_k,p_e>\in \mathbf{s}} p(e^{p_e}|k^{p_k}) && (4) \\ &= \sum_{<p_k,p_e>\in \mathbf{s}} \sum_{\mathbf{a}(p_k,p_e)} p(e^{p_e}, \mathbf{a}(p_k,p_e)|k^{p_k}) && (5) \end{aligned}$$

If we represent the phrase-to-phrase correspondences using the tag sequence of phrase and words composing phrase, The equation 5 can be rewritten as in equation 6 letting phrase match be represented by the tag sequence of phrases as well as words. In equation 6, $k_j^{p_k}$ is j-th phrase of k^{p_k}, and $t(k_j^{p_k})$ denotes the tag sequence of words composing phrase $k_j^{p_k}$. $|p_e|$ is the number of phrases in a phrase sequence p_e.

$$\begin{aligned} & p(e^{p_e}|k^{p_k}) \\ &= \sum_{a_0=0}^{|p_k|} \cdots \sum_{a_{|p_e|}}^{|p_k|} \prod_{i=1}^{p_e} p(t(e_i^{p_e})|t(k_{a_i}^{p_k}))p(e_i^{p_e}|k_{a_i}^{p_k}) \\ &= \prod_{i=1}^{|p_e|} \sum_{j=0}^{|p_k|} p(t(e_i^{p_e})|t(k_j^{p_k}))p(e_i^{p_e}|k_j^{p_k}) && (6) \end{aligned}$$

The likelihood of all alignable cases within bilingual phrase is defined as in equation 7, where $|e_i^{p_e}|$

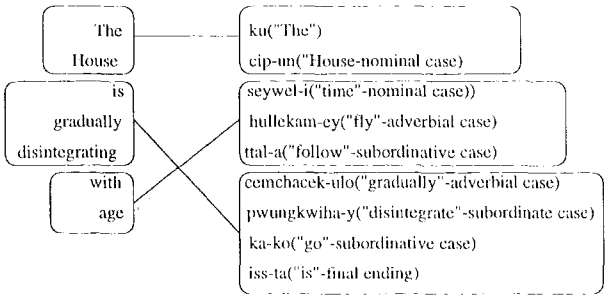

Figure 3: An example of Korean-English alignment at phrase level.

is the number of words in a phrase $e_i^{p_e}$ and $e_{ik}^{p_e}$ denotes k-th word of in a phrase $e_i^{p_e}$.

$$p(e_i^{p_e}|k_j^{p_k}) = \prod_{k=1}^{|e_i^{p_e}|} \sum_{l=1}^{|k_j^{p_k}|} p(e_{ik}^{p_e}|k_{jl}^{p_k}) \quad (7)$$

Figure 3 shows how the problem of word unit mismatch can be dealt with in the phrase level alignment.

In the example, e^{p_e} = (The house) (is gradually disintegrating) (with age), and $e_1^{p_e}$ = (The house), $e_{11}^{p_e}$ = The, $t(e_1^{p_e})$ = (determiner noun), $k_1^{p_k}$ = (ku cip-un), $k_{11}^{p_k}$ = ku, respectively.

2.4 Parameter reestimation

With the constraint that the sum over all alignments should be 1, the reestimation algorithm can be derived to give equation 8 for word translation probability and equation 10 for phrase correspondence probability. This process, when applied repeatedly, must give a locally optimal estimation of the parameters following the principle of the EM algorithm (Brown et al. 1993)(Dempster et al. 1977).

$p(e|k)_{<condition>}$ denotes the alignment candidates that satisfies $<$ *condition* $>$. For calculating $p(e|k)$, only constant t cases of alignments need to be considered in the proposed alignment algorithm because most alignment candidates have very low probability that they may be ignored.

$$p(e|k) = \frac{\text{expected number of } e \text{ given } k}{\text{total expected number of } e \text{ given } k} = \frac{\sum_{\mathbf{e},\mathbf{k}\in\mathbf{corpus}} c(e|k;\mathbf{e},\mathbf{k})}{\sum_{e}\sum_{\mathbf{e},\mathbf{k}\in\mathbf{corpus}} c(e|k;\mathbf{e},\mathbf{k})} \quad (8)$$

Let us call the expected number of times, that k matches with e in the corresponding sentence $\mathbf{k}$ and $\mathbf{e}$, the count of $\mathbf{e}$ given $\mathbf{k}$. By using the notation $c(e|k)$, the reestimation formula of $p(e|k)$ can be induced as equation 8 using EM method.

$$c(e|k;\mathbf{e},\mathbf{k}) = \frac{P(\mathbf{e}|\mathbf{k})_{<e=e_{ik}^{p_e},k=k_{jl}^{p_k}>}}{p(\mathbf{e}|\mathbf{k})} \quad (9)$$

When we denote $c(t_e|t_k)$ the expected number of times that a tag sequence of English phrase corresponds to a tag sequence of Korean phrase as in equation 11. Then the reestimation algorithm of $p(t_e|t_k)$ is given as in equation 10.

$$p(t_e|t_k) = \frac{\text{expected number of } t_e \text{ given } t_k}{\text{total expected number of } t_e \text{ given } t_k} = \frac{\sum_{\mathbf{e},\mathbf{k}\in\mathbf{corpus}} c(t_e|t_k;\mathbf{e},\mathbf{k})}{\sum_{t_e}\sum_{\mathbf{e},\mathbf{k}\in\mathbf{corpus}} c(t_e|t_k;\mathbf{e},\mathbf{k})} \quad (10)$$

$$c(t_e|t_k;\mathbf{e},\mathbf{k}) = \frac{P(\mathbf{e}|\mathbf{k})_{<t_e=te_i^{p_e},t_k=k_j^{p_k}>}}{p(\mathbf{e}|\mathbf{k})} \quad (11)$$

For the extended method of phrase alignment, the base model is an intermediate stage for the estimation of word-to-word probabilities. The phrase-to-phrase probabilities are reestimated upon the initial values of word-to-word probabilties.

2.5 Alignment algorithm

The alignment process of generating Korean phrases and selecting their matching phrases in English can be formulated around the principle of dynamic programming. The probability value defined in equation 6 and 7 is used to compute matching probability of $p(k_{i,a})$ and $p(e_{j,b})$.

$p(e_{j,b})$ stand for the phrase composed of b number of words from j-th word in a sentence. ζ_i is used to keep the selected phrase sequence up to i-th word and φ_i denotes its score. N and M are number of words of Korean sentence and English sentence respectively. The constant value L is defined as maximum number of words which consist of a phrase.

Initialization

$$\varphi_0 = 0$$

Recursion

$$\varphi_i = \max_{\substack{1\le i\le N\\ 1\le a\le L\\ 1\le b\le L}} [\varphi_{i-a} + \log p(k_{i,a}, e_{j,b})]$$

$$\zeta_i = (j,a,b) = \arg\max_{\substack{1\le i\le N\\ 1\le a\le L\\ 1\le b\le L}} [\varphi_{i-a} + \log p(k_{i,a}, e_{j,b})]$$

Path backtracking

$$\text{optimal path} = (\zeta_0, ...\zeta_{h_{m-1}}, \zeta_{h_m}, ..., \zeta_N)$$

$$h_{m-1} = h_m - a(\zeta_{h_m}), \text{where } a(\zeta_{h_m}) \text{ is } a \text{ in } \zeta_{h_m} = (j,a,b)$$

Although the alignment algorithm described above with the complexity of $O(L^2MN)$ is simple and efficient, this algorithm has the limitation caused by the assumption of dynamic programming. The dynamic programming in the context of alignment assumes that the previous

selections do not interfere with the future decisions. The alignment decision, however, may depend on the previous matches to the extent that the results from dynamic programming may not be sufficiently accurate. One popular solution is to maintain upper t-best cases instead of just one as following where max-t denotes the t-th max candidate.

$$\varphi_i(t) = \max\text{-}t_{\substack{1\le t'\le T,1\le j\le N\\ 1\le a\le L,1\le b\le L}} [\varphi_{i-a}(t') + \log p(k_{i,a}, e_{j,b})]$$

$$\zeta_i(t) = (j,a,b)$$

$$= \arg \max\text{-}t_{\substack{1\le t'\le T,1\le j\le N\\ 1\le a\le L,1\le b\le L}} [\varphi_{i-a}(t') + \log p(k_{i,a}, e_{j,b})]$$

As a result, the running complexity of the proposed algorithm becomes $O(TL^2MN)$. Taking T and L as constants, the order of complexity becomes $O(MN)$.

As another method to relax the problem of decision dependency on the previous matches, preemptive scheme to find max matching of phrase $k_{i,a}$ is adopted. In the preemptive alignment, the previous selection can be rematched with the better selection found by later decision.

In following algorithm, $\varrho(k_{i,a}, n)$ denote $e_{j,b}$ which has n-th highest matching value with Korean phrase $k_{i,a}$ among all possible matching Korean phrase and $\nu(k_{i,a}, n)$ carry the weight for the matching. $\xi_{j,b}$ indicate matched Korean phrase with $e_{j,b}$ in current status and $\vartheta_{j,b}$ denote their matching weight. The established matching in previous stage can be changed when another matching, which has higer matching weight, is identified in this algorithm.

Initialization

$$\varrho(k_{i,a}, n) = (j,b)$$

$$\vartheta_{j,b} = 0\ , (1 \le j \le N, 1 \le b \le L)$$

$$\nu(k_{i,a}, n) = \frac{p(k_{i,a}, e_{j,b})}{\sum_{j=1}^{N}\sum_{b=1}^{L} p(k_{i,a}, e_{j,b})}$$

Preemptive selection

$n = 0$
$(j,b) = \varrho(k_{i,a}, n)$
repeat
  $if(\nu(k_{i,a}, n) > \vartheta_{j,b})$
    $\vartheta_{j,b} = \nu(k_{i,a}, n)$
    $k'_{i,a} = \xi_{j,b}, \xi_{j,b} = k_{i,a}, k_{i,a} = k'_{i,a}$
  else
    $n = n + 1, (j,b) = \varrho(k_{i,a}, n)$
until $\vartheta_{j,b}$ is 0

Table 2: The content of training corpus (English:words, Korean:word-phrases)

Source	English	Korean
middle-school textbook	46,400	34,800
high-school textbook	153,300	106,400
other books	54,400	37,100
total	254,100	178,300

Although the proposed algorithm can not cover all possible alignment cases, the proposed algorithm produces resonably accurate alignment results efficiently as is demonstrated in the following section.

2.6 Experiments

The total training corpus for our experiments consists of 254,100 English words and 178,300 Korean word-phrases. The content of training corpus is summarized in table 2.

A HMM Part-of-Speech tagger is used to tag words before alignments. An accurate HMM designed by the authors for Korean sentences taking into account the fact that a Korean sentence is a sequence of word-phrases is used (Shin et al. 95). The Penn Treebank POS tagset that is composed of 48 tags and 52 Korean tagset is used in the tagging. The errors that is generated by morphological analysis and tagging cause many of the alignment errors.

To avoid the noise due to the insufficient bilingual sentences, we adopted two significance filtering methods that were introduced by Wu and Xia (1994). First, the Korean sentences consisting of words with more than 5 occurrences in the corpus are considered in the experiment. Second, we selected the English words that accounts for the top 0.80 of the translation probability density given a Korean word.

When we selected 200 sentence pairs randomly and manually tested aligned results, we obtained 68.7% precision at the phrase level and 89.2% precision of bilingual dictionary induced from the alignment. The table 3 and 4 illustrate the bilingual knowledge acquired from the aligned results. The information in table 4 is the unique product of phrase-level alignment.

3 Conclusion

With the alignment of Korean-English sentences, the most serious problem, that is seldom found at Indo-European language pairs, is how to overcome the differences of word unit and word order. The proposed method is an extension of word level alignment and solves the problems of word unit mismatch and word order through phrase level alignment. We have also described several alternatives of alignment and parameter estimation.

Table 3: Examples of result for word translation probability.

Korean word	English word	probability
yengli	clever	0.616331
yengli	smart	0.238197
yengli	cleverness	0.145472
kion	degrees	0.279992
kion	temperatures	0.248706
kion	centigrade	0.131713
kion	increase	0.130894
kion	would	0.108766

Table 4: Examples of phrase-level bilingual dictionary results

Korean phrase	English phrase
wuli uy chwuswu kamsacel	our thanksgiving day
cy kwansim i iss	interested in
maywu kupkyek ha ko tto wuihem	very fast and dangerous
moscianh key wuihem	dangerous as anything else

It produces more accurate bilingual dictionary than the method using only word correspondence information. Moreover, we can extract phrase-level information from the results of phrase level alignment. Also in the proposed method, the whole process of generating phrase units and finding matching phrases, is done mechanically without human intervention. One negative aspect is that the method requires large amount of training corpus for the saturated estimation of the model though larger data will increase the accuracy of the performance.

The proposed method may well be applied to other language pairs of similar structures as well as dissimilar ones. Since the results from the method are richer with linguistic information, other applications such as machine translation and multilingual information retrieval are promising research areas.

References

Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra. 1995. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39-73.

A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39:1-38,1977.

Dekai Wu, Xuanyin Xia. 1994. Learning an Enlish-Chinese lexicon from a parallel corpus. In *Proceedings of AMTA-94*, 206-213. Columbia.

Grammarless extraction of phrasal translation examples from parellel corpra 1995, In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, 354-371. Leuven, Begiun.

Frank A. Smadja. 1992. How to compile a bilingual collocational lexicon automatically. In *AAA-92 Workshop on Statistically-Based NLP Techniques*, 65-71, San Jose, CA.

Judith Klavans, Evelyne Tzoukermann. 1990. The bicord system. In *Proceedings of COLING-90*, 174-179. Helsinki, Finland.

Jung H. Shin, Young S. Han, Young C. Park, Key-Sun. Choi. 1995. A HMM Part-of-Speech Tagger for Korean with wordpharsal Relations. In Proceedings of Recent Advances in Natural Language Processing.

Peter F. Brown, Stephen A. Deela Pietra, Vincent J. Della Pietra, Robert L. Mercer.1991. Word Sense disambiguation using statistical methods. In *Proceedings of 29th Annual Meeting of ACL*, Berkeley CA.

Peter F. Brown, Stephen A. Deela Pietra, Vincent J. Della Pietra, Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263-311.