Improving Lexical Mapping Model of English-Korean Bitext Using Structural Features

Seonho Kim† and **Juntae Yoon**‡ and **Mansuk Song**†
†Dept. of Computer Science, Yonsei University,
Seoul 120-749, Korea
shkim,mssong@december.yonsei.ac.kr
‡NLP Lab., Daumsoft Inc., Kangnam-gu, Samsung-dong,
Yungjeon Bldg., Seoul 135-728, Korea
jtyoon@daumsoft.com

Abstract

The problem of finding lexical alignments for given sentence pairs is computationally expensive. Furthermore, it is much difficult to find lexical alignments between Korean and English since they have considerably different syntactic structures and the coverage of word-for-word correspondences is low. This paper presents a method for extracting structural features which can reduce mapping space by allowing only probable alignments. We describe how the features improve the performance of the lexical alignment model. The structural features provide the information for the correspondences of parts-of-speech (POS) sequences which are useful in translation. Based on maximum entropy (ME) concept, the structural features are incrementally selected, which are later embedded in the lexical alignment model. It turns out that the features help get better lexical alignments of Korean and English by offering linguistic knowledge.

1 Introduction

Aligned bitexts are useful for the derivation of bilingual lexical resources which are used for machine translation and cross languages information retrieval. Thus, a lot of approaches have been suggested to find sets of corresponding word tokens (Brown et al., 1993; Berger et al., 1996; Melamed, 1997), phrase (Shin et al., 1996), noun phrase (Kupiec, 1993), and collocation (Smadja et al., 1996) in a bitext.

Some works have used lexical association measures for finding word correspondences (Gale and Church, 1991; Fung and Church, 1994). However, the association measures can be misled in cases where a word in a source language frequently co-occurs with more than one word in a target language or in cases of indirect association[1](Melamed, 1997).

In other works, iterative parameter re-estimation techniques based on IBM model 1∼5 [2] have been employed (Brown et al., 1993). They were usually incorporated in the EM algorithm (Brown et al., 1993; Kupiec, 1993; Tillmann and Ney, 2000; Och et al., 2000). However, we are often faced with some difficulties as follows, when the IBM model-based approaches are directly applied to the alignment, especially on bitext involving a less closely related language pair.

1. It needs excessive iteration time for parameter estimation and high decoding complexity. Thus most systems assumed one-to-one correspondence to reduce computational complexity. However, word sequences are not translated literally word for word. For example, in cases of collocations, compound nouns, and ambiguous words with different meaning dependent on the context, they require phrase-level correspondences.
2. Most systems use little or no linguistic knowledge to structure the underlying models. The distortion probability and the fertility probability for finding word correspondences is a weak description for word order change between languages and 1:n mapping modeling. As the result, lots of ungrammatical sentences and too many parameters to be estimated are allowed. In addition, many words are aligned to the empty word due to the overfitting problem (Och et al., 2000).
3. In order to estimate parameters properly, it requires a very large volume of bilingual aligned text.

In this paper, we use structural correspondences to

[1] Suppose that u_k and v_k are indeed mutual translation and u_k and u_{k+1} often co-occur in text. Then v_k and u_{k+1} will also co-occur more than expected by chance , which is represented as **indirect association**.

[2]
$$p(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{i=1}^{l} \mathbf{n}(\phi_i|e_i) \prod_{j=1}^{m} \mathbf{t}(f_j|e_{a_j}) \mathbf{d}(j|a_j, m, l)$$

n is the fertility probability that an English word generates n French words, **t** is the word alignment probability that the English word e generates the French word f, and **d** is the distortion probability that an English word in a certain position will generate a French word in a certain position.

English	Korean
can/MD	*eul*/ENTR1 *su*/NNDE *iss*/AJ *da*/ENTE
World/NNP Cup/NNP	*weoldeukeop*/NNIN1
to/TO take/VBP off/RP	*iryugha*/VB *neun*/ENTR1
MD:modal auxiliary, NNP:proper noun, VBP:verb, RP:particle NNDE:dependent noun, AJ:adjective ENTE: final ending, NNIN1:proper noun, VB:verb, ENTR1:adnominal ending	

Table 1: Lexical differences between English and Korean

overcome the above problem in bilingual text alignment.

2 The problems of lexical alignment in Korean and English bitext

Although one-to-one correspondence assumption has been shown to give highly accurate results in closely related language pair (Melamed, 1997), it does not fit in structurally different language pair such as Korean and English. According to Shin et al. (1996), the coverage of one-to-one correspondence in Korean and English bitext is approximately 34% and many-to-many mappings exceed 40% in lexical alignments. Besides, Korean and English show much difference in terms of unit of mutual translation as shown in Table 1.

Thus, we use structural or linguistic information to find accurate lexical alignments of bitext. In general, structural information can be represented by the form of bilingual correspondence of phrase structure (Watanabe et al., 2000) or dependency structure (Yamamoto, 2000).

To find these structural correspondences between a language pair, unification-based grammar (Matsumoto, 1993) or bilingual grammar (Wu, 1997) can be used. That is, if we try to find structural match of bitext, syntactic analysis of each text should be done first. However, syntactic analysis of each text(or bitext) is also a hard and computationally complicated problem.

In this paper, we propose POS sequence feature. We consider correspondences between POS tag sequences of bitext as the structural information which is important in determining structures and reducing unnecessary parameters of a statistical alignment model. We induce the correspondences(structural features) by a feature selection method based on the ME framework. Finally, the structural features are embedded in a lexical alignment model of English and Korean bitext. Our model offers the following advantages:

1. It can overcome the limitation of word-for-word correspondences.
2. The model can take advantage of the explicit introduction of some knowledge about the language. Therefore, it can reduce a lot of parameters in statistical machine translation by eliminating syntactically unlikely alignments. The lexical alignment model iterates only over a subset of probable alignments.
3. It is possible to estimate the probability of lexical alignment in a relatively small size of corpora.
4. The structural information is helpful in the construction of a bilingual grammar.

3 Overview of the model

In general, there exist constaints among POS sequences mapping when aligning bitext. For instance, in many cases a noun in Korean is translated to a noun or a noun preceded by an article in English. This POS constraint would be useful information in alignment of a closely related pair as well as a less closely related language pair.

Some approaches design an alignment algorithm that maximizes the number of matching POS in aligned segments (Papageorgiou et al., 1994). We also assume that the mapping information of POS sequences of a language pair is useful in a model of statistical machine translation.

If there are similarities between corresponding POS sequences in bitext, the structural feature would be easily computed or identified. However, a POS sequence in English often correspond to a totally different POS sequence in Korean as shown in the following example:

can/MD $\rightarrow \sim$ *eul*/ENTR1 *su*/NNDE1 *'iss*/AJMA *da*/ENTE

It is caused by the discrepancy between two languages.

Korean is an agglutinative language. A sentence in Korean consists of a series of syntactic units called *eojeol*. An eojeol delimited by whitespace is often composed of a content word and function words. Tense markers, clausal connectives, particles and so forth are contained in an eojeol. Thus, one or more words in English often correspond to an eojeol, i.e. a couple of morphemes. For instance, a phrase, 'to the school', in English corresponds to an eojeol '학교로(*haggyo-ro*, school-to)' in Korean.

Thus, we need a method for mapping POS sequences of bitext. In this paper, the correspondences of POS tag sequences are obtained by automatic feature selection based on the ME framework. Here, a feature is defined as a correspondence of POS tag sequences in bitext.

The outline of finding correspondences of POS tag sequences is as follows: First, our model starts with initial features obtained by a supervision step. The initial features are extracted from a small portion of bitext, which of weights are trained by the IIS algorithm (Berger et al., 1996; Pietra et al., 1997). These are called *initial active features*.

In the next step, a feature pool is constructed from training data. At this time, only features giving a large gain to the model are selected. The final output of feature selection is the set of *active features* and the correspondences of POS tag sequences that are represented with conditional probabilities.

The resulted features are used for the parameters of bilingual text alignment. In the process, we look at the words to ensure a correct alignment. That is, the POS sequences are in fact encoding particular words.

The underlying process is as follows:

Input: a set **L** of POS-labeled sentence pairs.

1. Make a set $\mathcal{F}$ of correspondences of tag sequences, (t_e, t_k) from a small portion of **L** by hand.
2. Set $\mathcal{F}$ into a set of initial active features, $\mathcal{A}$.
3. Compute the weights of the initial active features λ using IIS algorithm.
4. Create a feature pool $\mathcal{P}$ which is a set of possible combinations of tag sequences from sentence pairs.
5. Filter $\mathcal{P}$ using frequency counts and similarity with $\mathcal{A}$.
6. Compute the approximate gains of the features in $\mathcal{P}$.
7. Select features($\mathcal{N}$) with large gain values, and add $\mathcal{A}$.
8. Compute the lexical alignment of bitext using $p(t_k|t_e)$ *where* $(t_e, t_k) \in \mathcal{A}$.

---

4 Feature Selection

As mentioned above, features which represent mappings of POS sequences in two languages are automatically learned from bitext. The learning consists of two steps: (1) supervised step and (2) unsupervised feature selection.

In the supervised step, we manually aligned English and Korean texts. From the manually aligned text, we consturct a feature pool which has initial POS sequence correspondence. In the unsupervised feature selection, the POS sequence correspondeces are added to the feature pool.

4.1 Supervision Step

In the supervision step, a small portion of bitext is tagged using Brill's tagger (Brill, 1995) and 'MORANY' (Yoon et. al., 1999), each of which is for English and Korean tagging respectively. We manually aligned each sentence pair in the bitext and collected their correspondences of tag sequences. For simplicity, we adjusted some part of Brill's tag set.

First of all, we classified sentential patterns of English and Korean. Then, we aligned English and Korean sentence pairs on the basis of the patterns. Also, we made tag sequence construction rules with respect to each language by analysis of the cases where the words of one unit are separated and adjacent. The rules are used when making a feature pool.

After collecting the correspondences of POS tag sequences, we used them as a set of initial active features, $\mathcal{A}$.

4.2 Construction of a Feature Pool

In this chapter, we describe how a feature pool is constructed. Our task is to construct a probability model p that produces a corresponding Korean tag sequence $\mathcal{T}_\mathbf{k}$ for a given English tag sequence $\mathcal{T}_\mathbf{e}$. As features to represent the model, we use co-occurrence information of POS tag sequences.

Let $\mathcal{T}_S$ be all possible correspondences of tag sequences for a specific sentence pair, S. We then define a feature function (or a feature) as follows:

$$f_{t_e,t_k}(x,y) = \begin{cases} 1 & x = t_e \;\&\; y = t_k \;\& \\ & pair(t_e, t_k) \in \mathcal{T}_S \\ 0 & otherwise \end{cases}$$

A feature f_{t_e,t_k}, which indicates co-occurrence between tags appearing in $\mathcal{T}_S$, expresses information for predicting that an English POS tag sequence t_e maps into a Korean POS tag sequence t_k.

In order to make a feature pool, given a sentence pair, we first construct all possible combinations of English POS tag sequences and Korean POS tag sequences. Among them, only features f_{t_e,t_k} that satisfy the following two conditions are added into the feature pool $\mathcal{P}$.

- $count(f_{t_e,t_k}) \geq 15$
- there exist t_{k_x} such that $(t_e, t_{k_x}) \in \mathcal{A}$ and the similarity value (simply counting of same tag) between t_k and t_{kx} is greater than 0.6

4.3 Feature selection

Since the set of $\mathcal{P}$ is too vast, a feature selection process is needed to select useful features. For this, each feature is evaluated according to how much they contribute to the likelihood of training data, which is called *gain*.

Before explaining feature gain and a process of feature selection, we will give a brief introduction of ME. In ME, a feature gives information to a probability model and it has a weight.

Thus, the model is constrained by features we defined. In general weights of the features are trained by the improved iterative scaling (IIS) algorithm that minimizes the Kullback-Leibler divergence be-

tween the model and the empirical distribution of the training data.

In fact, our model reduces to a simple type of probability model that can be derived simply from a ratio counts since the features do not overlap.

Let $p_{\mathcal{A}}$ be the optimal model constrained by a set of initial active features $\mathcal{A}$, then it can be represented in (1).

$$p_{\mathcal{A}}(y|x) = \frac{1}{Z_{\mathcal{A}}} p(y|x) e^{\sum_i \lambda_i f_i(x,y)} \quad (1)$$
$$Z_{\mathcal{A}}(x) = \sum_y p(y|x) e^{\sum_i \lambda_i f_i(x,y)}$$

The weights (λ) of active features are computed by the IIS before feature selection.

Let $\mathcal{A}f_i$ be $\mathcal{A} \cup f_i$, and $p_{\mathcal{A}f_i}$ be the optimal model in the space of probability distribution after adding feature f_i. The model $p_{\mathcal{A}f_i}$ contains another parameter α, in addition to the parameters given by active features, which is a weight for the feature f_i. In order to make it tractable to compute feature selection, we assume that the addition of a feature f_i affects only the single parameter α. The only parameter which distinguishes models of (2) is α.

$$p_{\mathcal{A}f_i} = \frac{1}{Z_\alpha(x)} p_{\mathcal{A}}(y|x) e^{\alpha f_i(x,y)} \quad (2)$$
$$Z_\alpha(x) = \sum_y p_{\mathcal{A}}(y|x) e^{\alpha f_i(x,y)}$$

The improvement (gain) of a model after adding a single feature f_i can be estimated by measuring difference of maximum log-likelihood between $L(p_{\mathcal{A}f_i})$ and $L(p_{\mathcal{A}})$. We denote the gain for feature f_i by $\Delta(\mathcal{A}f_i)$, which is represented in (3).

$$\begin{aligned} \Delta(\mathcal{A}f_i) &\equiv max_\alpha G_{\mathcal{A}f_i}(\alpha) \quad (3) \\ G_{\mathcal{A}f_i}(\alpha) &\equiv L(p_{\mathcal{A}f_i}) - L(p_{\mathcal{A}}) \\ &= \log \frac{p_{\mathcal{A}f_i}}{p_{\mathcal{A}}} \end{aligned}$$

The following algorithm is used to compute the gain of the model with respect to f_i. We adopted Berger's method for computing gains (Berger et al., 1996). For the details, the reader is referred to (Berger et al., 1996; Pietra et al., 1997).

1. Let $r = \begin{cases} 1 & if\ \tilde{p}(f_i) \le p_{\mathcal{A}}(f_i) \\ -1 & otherwise \end{cases}$
2. Set $\alpha_0 = 0$
3. Repeat the following until $G_{\mathcal{A}f_i}(\alpha_n)$ has converged :
   Compute α_{n+1} from α_n using
   $\alpha_{n+1} = \alpha_n + \frac{1}{r} \log\left(1 - \frac{1}{r} \frac{G'_{\mathcal{A}f_i}(\alpha_n)}{G''_{\mathcal{A}f_i}(\alpha_n)}\right)$
   Compute $G_{\mathcal{A}f_i}(\alpha_{n+1})$ using
   $G_{\mathcal{A}f_i}(\alpha) = -\sum_x \tilde{p}(x) \log Z_\alpha(x) + \alpha \tilde{p}(f_i)$,
   $G'_{\mathcal{A}f_i}(\alpha) = \tilde{p}(f_i) - \sum_x \tilde{p}(x) M(x)$,
   $G''_{\mathcal{A}f_i}(\alpha) = -\sum_x \tilde{p}(x) p^\alpha_{\mathcal{A}f_i}((f_i - M(x))^2|x)$
   where $\alpha = \alpha_{n+1}$, $\mathcal{A}f_i = \mathcal{A} \cup f_i$,
   $M(x) \equiv p^\alpha_{\mathcal{A}f_i}(f_i|x)$,
   $p^\alpha_{\mathcal{A}f_i}(f_i|x) \equiv \sum_y p^\alpha_{\mathcal{A}f_i}(y|x) f_i(x,y)$
4. Set $\sim \Delta L(\mathcal{A}f_i) \leftarrow G_{\mathcal{A}f_i}(\alpha_n)$

This algorithm is iteratively computed using Netwon's method. With the gain value, we can recognize importance of a feature, i.e. how much the feature accords with the model. As a result, we can select useful features with high gain values and obtain their conditional probabilities. Put another way, we can get the correspondence probabilities of POS tag sequences.

5 lexical alignment

In this section, we describe how the correspondence probabilities between bilingual POS sequences is embedded in lexical alignment. In Korean-English machine translation, a translation system finds the corresponding sentence **e** given a Korean sentence **k**.

The fundamental equation of machine translation can be represented in (4), where the random variables **K** and **E** are a Korean sentence and a English sentence making up a translation, and the random variable **A** is an alignment between them. The equation is related with a language model probability, $P(\mathbf{e})$, a translation model probability $P(\mathbf{k}|\mathbf{e})$. In this paper, we are interested in estimating the translation model probability $P(\mathbf{k}|\mathbf{e})$.

$$P(\mathbf{e}|\mathbf{k}) = \frac{P(\mathbf{e})P(\mathbf{k}|\mathbf{e})}{P(\mathbf{k})}$$
$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} P(\mathbf{e})P(\mathbf{k}|\mathbf{e}) \quad (4)$$

In this work, we modified the IBM 1 model using the structural information for estimating translation probabilities. The translation probability of a specific sentence pair, **k** and **e** is

$$P(\mathbf{k}|\mathbf{e}) = \epsilon \prod_{j=1}^{m} \sum_{i=1}^{l} t(k_{p_j}|e_{p_i}) P(C_{k_{p_j}}|C_{e_{p_i}}) \quad (5)$$

In (5), an English sentence **e** has l possible phrases, $e_{p_1}e_{p_2}...e_{p_l}$, and a Korean sentence **k** has m possible phrases, $k_{p_1}k_{p_2}...k_{p_m}$ and the phrases of **e** have corresponding sequences of POS categories, $C_{e_{p_1}}C_{e_{p_2}}...C_{e_{p_l}}$, and the phrases of **k** have the sequences of POS categories $C_{k_{p_1}}C_{k_{p_2}}...C_{k_{p_m}}$. ϵ is

		English	Korean
Train	Sentences	30,000	
	Words/Eojeol	282,967	201,771
	Vocabulary	47,647	80,360
	Morph Types	52,197	57,800

Table 2: Training corpus size

a small, fixed normalizing number (Brown et al., 1993).

Note that the probability $P(C_{k_{p_j}}|C_{e_{p_i}})$ is precomputed by the feature selection process and only lexical alignment parameter $t(k_{p_j}|e_{p_i})$ can be estimated by the EM algorithm using sentence pairs of bitext, $(\mathbf{k^{(s)}}, \mathbf{e^{(s)}})$, $\mathbf{s} = 1, ..., \mathbf{S}$. According to the equation (4), we can eliminate the combination that $P(C_{k_{p_j}}|C_{e_{p_i}})$ is zero. Thus, the model iterates only over a subset of probable alignments.

To estimate probability $t(k_p|e_p)$, the expected number of times(fractional counts) that the phrase e_p connects to k_p in the translation $(\mathbf{k}|\mathbf{e})$ is used. The count denoted by $c(k_p|e_p; \mathbf{k}, \mathbf{e})$ can be expressed in (7) using (6) and translation probabilities are reestimated by (8).

$$P(\mathbf{a}|\mathbf{e}, \mathbf{k}) = \frac{P(\mathbf{k}, \mathbf{a}|\mathbf{e})}{P(\mathbf{k}|\mathbf{e})} = \frac{P(\mathbf{k}, \mathbf{a}|\mathbf{e})}{\sum_{\mathbf{a}} P(\mathbf{k}, \mathbf{a}|\mathbf{e})} \quad (6)$$

$$c(k_p|e_p; \mathbf{k}, \mathbf{e}) = \mathcal{E} \sum_{j=1}^{m} \delta(k_p, k_{p_j}) \sum_{i=1}^{l} \delta(e_p, e_{p_i})$$

$$\mathcal{E} = \frac{t(k_p|e_p)P(C_{k_p}|C_{e_p})}{t(k_p|e_{p_1})P(C_{k_p}|C_{e_{p_1}}) + ... + t(k_p|e_{p_l})P(C_{k_p}|C_{e_{p_l}})} \quad (7)$$

$$t(k_p|e_p) = \frac{c(k_p|e_p; \mathbf{k}, \mathbf{e})}{\sum_{k_p} \sum_{s=1}^{S} c(k_p|e_p; \mathbf{k}^{(s)}, \mathbf{e}^{(s)})} \quad (8)$$

6 Experiments

We present results tested on English-Korean bitext that is extracted from the web site of 'Korea Times' and a magazine for English learning (Table 2).

We manually aligned 700 POS-tagged sentence pairs to obtain initial parameters for correspondences of tag sequences. As shown in Table 3, the coverage of word-for-word correspondences in English-Korean bitext was only 31.2%.

As a result, 1,483 correspondences of tag sequences were collected from the manually aligned bitext (Table 4).

The correspondences were used as initial active features and the weights of the initial active features were computed by IIS algorithm. Table 5 shows weights λ of the initial active features such that $f_{t_{BEP+JJ},t_k} \in \mathcal{A}$.

English words	Korean morphemes	Ratio
1	1	31.2
1	2	22.4
1	3	13.6
2	1	6.9
etc	etc	25.9

Table 3: The result of alignment unit

set	description	disjoint features
A	active features	1,483
P	filtered feature pool	8,147
N	selected new features	702

Table 4: Features(Correspondences of Tag Sequence)

Through the process of feature pool construction, 8,147 features(tag sequence correspondences) were selected from the 23,000 sentence pairs. Since we used the filtering method and tag sequence construction rules, the size of the feature pool was not large.

In the feature selection step, we chose useful features with the gain threshold of 0.008. As a result of feature selection, we obtained the probability model that given a specific English POS tag sequence, a corresponding Korean POS tag sequence happens to occur.

Table 6 shows some examples of conditional probabilities. The table shows that the determiner of English is generally translated into NULL or adnominal word in Korean. The conditional probabilities regarding the correspondences of POS tag sequences were used as known parameters of the lexical alignment model.

Effect of Smoothing

As mentioned before, in previous IBM-based models, many words are aligned to the empty word and rare words are mis-aligned. Thus, we tested if the structural features are effective in smoothing. For this, the accuracy of lexical correspondences was evaluated both on low frequency words and high fre-

t_e	t_k	$p(t_k\|t_e)$
DT+NN	NNIN2	0.524131
DT+NN	ANDE+NNIN2	0.15161
DT+NN	ANNU+NNDE2	0.091036
DT+NN	NNIN2+PPCA1	0.063515
DT+NN	NNIN2+NNIN2	0.058322
DT+NN	NNIN2+PPAU	0.05768
DT+NN	ADCO	0.049622
etc	etc	

Table 6: Examples of conditional probability

f_{t_e,t_k}	λ_i	p(t_k|t_e)	e	k
$f_{BEP+JJ,VBMA+ENCO3+AX+ENTE}$	10.1369	0.4247	are+prepared	junbidoi+eo+iss+da
$f_{BEP+JJ,VBMA}$	8.8520	0.1180	is+careful	ju'yiha
$f_{BEP+JJ,AJMA}$	8.6787	0.0996	am+healthy	geongangha
$f_{BEP+JJ,AJMA+ENTE}$	8.2628	0.0655	is+new	syaelob+da
$f_{BEP+JJ,VBMA+ENTE}$	7.2379	0.0236	am+sure	hwagsinha+bnida
$f_{BEP+JJ,NNIN2+CO}$	7.1372	0.0210	am+rich	buja+i
$f_{BEP+JJ,NNIN2+CO+VBMA}$	6.9909	0.0183	is+selfish	igijeog+i+doi
$f_{BEP+JJ,NNIN2+PPCA1+VBMA+ENTE}$	6.8402	0.0157	is+patriotic	'aegugja+ga+doi+da
$f_{BEP+JJ,NNIN2+CO+ENTE}$	6.8308	0.0156	is+reasonable	habligeog+i+da
$f_{BEP+JJ,NNIN2+PPCA2+AX+ENTE}$	6.4256	0.0105	is+rephrehensible	binanbad+eul+manha+da
$f_{BEP+JJ,NNIN2+PPCA1+VBMA}$	6.4250	0.0105	am+helpful	doum+i+doi+da

BE:be verb(present tense), DT:determiner, JJ:adjective(ordinal), NN:common noun, RB:adverb
AJMA:adjective, VBMA:verb, AX:auxilary verb ENCO3:auxilary ending, ENTE:final ending
ANCO:configurative adnominal, ANDE: demonstrative adnominal , ANNU:numeral adnominal
NNIN2:common noun, NNDE2:unit-dependent noun, CO:copular
PPCA1:nominative postposition, PPCA2:accusative postposition, PPAU:auxilary postposition

Table 5: Examples of active features

frequency	English Words (Total)	Korean Eojeols	Accuracy
4	100(2149)	100(3317)	70.3%
50∼300	100(712)	100(486)	78.9%

Table 7: Evaluation Vocabulary

quency words.

Table 7 shows the accuracy of some results obtained by lexical alignment. In the training bitext, English 2149 words and Korean 3317 *eojeol*s with low frequency 4 and English 712 words and Korean 486 *eojeol*s with high frequency (50∼300) were found. For an evaluation, 100 words and 100 eojeols were selected out of them. The alignments of the words (*eojeol*s) were evaluated. As a result, it turns out that the structural features are effective in rare words.

Effect of reducing a parameter space

Another advantage of the use of structural features is to reduce parameter space of the lexical alignment model. To show the impacts on the size of the parameter space reduced by our model, we compared the results from our model with those from the IBM 1-model presented by Brown et al. (1993) on 100 sentence pairs. The number of parameters obtained means the counts of possible lexical mappings.

As shown in Table 8, the number of prameters were drastically reduced in our model. Considering the complexity of another models (IBM 2∼5) it is obvious that they have more parameters.

Effect of improving the results of lexical alignments

For evaluating the efficacy of the structural features, we compared the results with IBM model 1.

	the number of parameters
IBM 1	14,776
Our Model	1,344

Table 8: Problem Space

	Accuracy
IBM 1	59.8
Our Model	73.7

Table 9: Accuracy of n:1 and 1:1 alignments

As described before, only n(English):1 and 1:1 mapping are possible in IBM model since one-to-one correspondence is assumed. Thus, we selected only n:1 and 1:1 alignments out of the alignment results. For comparison, we investigated the alignments of the English 100 words with high frequency, which were explained above.

Table 9 shows the accuracy of lexical alignments. It is shown that the structural features have an effect on the alignment even though the amount of data investigated is small.

However, the overall accuracy of the alignment is somewhat low, which is mainly due to the small size of training samples. Table 10 shows some results of mutual translation. We see a considerable improvement when allowing for structural features (correspondences of POS tag sequences) in lexical alignments.

Error Analysis

Except the errors by the incorrect parameter estimation, most errors of correspondences of POS tag sequences are caused by POS tagging errors. In addition, the correspondences of adverbs turn out to be

Korean	English	Probability
jeongbu	government	0.7312
jeongbu	the government	0.2012
jeongbu	republic	0.0328
jeongbu	authorities	0.0171
jeongbu	officials	0.0119
jeongbu	Chinese	0.0041

Table 10: Examples of alignment

sometimes erroneous, which is due to the fact that the position of adverb can be moved quite free.

7 Conclusion

Because of the considerable difference between English and Korean, computation cost of lexical alignment is very high. One solution of the problem is to provide mapping information of syntactic structure between the two languages.

In this paper, we defined the structural feature as the correspondence of POS tag sequences and presented a method for extraction of structural features for Korean-English bilingual alignment. Firstly, the initial active features were collected from a small size of manually aligned bitext, which are trained by IIS which is a training algorithm for ME. Secondly, extracted from training data, the features giving a large gain were added to the set of active features.

Furthermore, the features extracted were tested for lexical alignment of bitext. The experiment showed that the features are helpful for reducing the mapping space in alignment. We expect that the alignment can be more accurate and efficient by combining the structural features with translation lexicon in the future.

References

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39-73.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543-565.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263-311.

A. P. Dempster, N. M. Laird and D. B. Rubin. 1976. Maximum likelihood from incomplete data via the EM algorithm. *The Royal Statistics Society*, 39(B) 205-237.

Pascale Fung and Kenneth W. Church. 1994. *Kvec: A New Approach for Aligning Parallel Texts* In *Proceedings of COLING 94*, 1096-1102.

William A. Gale and Kenneth W. Church. 1991. Identifying Word Correspondance in Parallel Text In *Proceedings of the DARPA & NLP Workshop*

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19:75-102.

Frederick Jelinek. 1997. *Statistical Methods for Speech Recognition* MIT Press.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of ACL 31*, 17-22.

Yuji Matsumoto. 1993. Structural matching of parallel texts. In *Proceedings of the 31st Annual Meeting of the ACL*, 23-30.

I. Dan Melamed. 1997. A word-to-word model of translation equivalence. In *Proceedings of ACL 35/EACL 8*, 16-23.

Franz Josef Och and Hermann Ney. 2000. Improving Statistical Alignment Models. In *Proceedings of ACL 38*, 440-447.

H. Papageorgiou, L. Cranias and S. Piperidis. 1994. Automatic Alignment in Parallel Corpora. In *Proceedings of ACL 32 (Student Session)*.

Stephen A. Della Pietra, Vincent J. Della Pietra, John D. Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380-393.

Kengo Sato 1998. Maximum Entropy Model Learning of the Translation Rules. In *Proceedings of ACL 36/COLING*, 1171-1175.

Jung H. Shin, Young S. Han, and Key-Sun Choi. 1996. Bilingual knowledge acquisition from Korean-English parallel corpus using alignment method. In *Proceedings of COLING 96*.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1-38.

Christoph Tillmann and Hermann Ney. 2000 Word Re-ordering and DP-based Search in Statistical Machine Translation. In *Proceedings of COLING 2000*.

Ye-Yi Wang and Alex Waibel. 1998. Modeling with structures in machine translation. In *Proceedings of ACL 36/COLING*

Hideo Watanabe, Sadao Kurohashi, and Eiji Aramaki. 2000. Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation. In *Proceedings of COLING 2000*.

Dekai Wu 1997. Stochastic Inversion Transduction Grammar and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23-3, pp. 377-403.

Kaoru Yamamoto and Yuji Matsumoto. 2000. Ac-

quisition of Phrase-level Bilingual Correspondence using Dependency Structure. In *Proceedings of COLING 2000*.

Juntae Yoon, Chunghee Lee, Seonho Kim, and Mansuk Song. 1999. Morphological Analyzer of Yonsei University., Morany: Morphological Analysis based on Large Lexical Database Extracted from Corpus. In *Proceedings of Hangule and Korean Information Processing Workshop*, pp.92-98.