Structural Feature Selection For English-Korean Statistical Machine Translation

Seonho Kim, Juntae Yoon, Mansuk Song
{pobi, jtyoon, mssong}@december.yonsei.ac.kr
Dept. of Computer Science,
Yonsei University, Seoul, Korea

Abstract

When aligning texts in very different languages such as Korean and English, structural features beyond word or phrase give useful information. In this paper, we present a method for selecting structural features of two languages, from which we construct a model that assigns the conditional probabilities to corresponding tag sequences in bilingual English-Korean corpora. For tag sequence mapping between two langauges, we first define a structural feature function which represents statistical properties of empirical distribution of a set of training samples. The system, based on maximum entropy concept, selects only features that produce high increases in log-likelihood of training samples. These structurally mapped features are more informative knowledge for statistical machine translation between English and Korean. Also, the information can help to reduce the parameter space of statistical alignment by eliminating syntactically unlikely alignments.

1 Introduction

Aligned texts have been used for derivation of bilingual dictionaries and terminology databases which are useful for machine translation and cross languages information retrieval. Thus, a lot of alignment techniques have been suggested at the sentence (Gale et al., 1993), phrase (Shin et al., 1996), noun phrase (Kupiec, 1993), word (Brown et al., 1993; Berger et al., 1996; Melamed, 1997), collocation (Smadja et al., 1996) and terminology level.

Some work has used lexical association measures for word alignments. However, the association measures could be misled since a word in a source language frequently co-occurs with more than one word in a target language. In other work, iterative re-estimation techniques have been employed. They were usually incorporated with the EM algorithm and dynamic programming. In that case, the probabilities of alignments usually served as parameters in a model of statistical machine translation.

In statistical machine translation, IBM 1∼5 models (Brown et al., 1993) based on the source-channel model have been widely used and revised for many language domains and applications. It has also shortcoming that it needs much iteration time for parameter estimation and high decoding complexity, however.

Much work has been done to overcome the problem. Wu (1996) adopted channels that eliminate syntactically unlikely alignments and Wang et al. (1998) presented a model based on structures of two languages. Tillmann et al. (1997) suggested the dynamic programming based search to select the best alignment and preprocessed bilingual texts to remove word order differences. Sato et al. (1998) and Och et al. (1998) proposed a model for learning translation rules with morphological information and word category in order to improve statistical translation.

Furthermore, many researches assumed one-to-one correspondence due to the complexity and computation time of statistical alignments. Although this assumption turned out to be useful for alignment of close languages such as English and French, it is not applicable to very different languages, in particular, Korean and English where there is rarely close correspondence in order at the word level. For such languages, even phrase level alignment, not to mention word alignment, does not gives good translation due to structural difference. Hence, structural features beyond word or phrase should be considered to get better translation between English and Korean. In addition, the construction of structural bilingual texts would be more informative for extracting linguistic knowledge.

In this paper, we suggest a method for structural mapping of bilingual language on the basis of the maximum entorpy and feature induction framework. Our model based on POS tag sequence mapping has two advantages: First, it can reduce a lot of parameters in statistical machine translation by eliminating syntactically unlikely alignments. Second, it can be used as a preprocessor for lexical alignments of bilingual corpora although it can be also exploited by itself for alignment. In this case, it would serve as the first step of alignment for reducing the parameter space.

2 Motivation

In order to devise parameters for statistical modeling of translation, we started our research from the IBM model which has been widely used by many researches. The IBM model is represented with the formula shown in (1)

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

(1)

Here, $\mathbf{n}$ is the fertility probability that an English word generates n French words, $\mathbf{t}$ is the alignment probability that the English word e generates the French word f, and $\mathbf{d}$ is the distortion probability that an English word in a certain position will generate a French word in a certain position. This formula is one of many ways in which $p(\mathbf{f}, \mathbf{a}|\mathbf{e})$ can be written as the product of a series of conditional probabilities.

In above model, the distortion probability is related with positional preference(word order). Since Korean is a free order language, the probability is not feasible in English-Korean translation.

Furthermore, the difference between two languages leads to the discordance between words that the one-to-one correspondence between words generally does not keep. The model (1), however, assumed that an English word can be connected with multiple French words, but that each French word is connected to exactly one English word including the empty word. In conclusion, many-to-many mappings are not allowed in this model.

According to our experiment, many-to-many mappings exceed 40% in English and Korean lexical alignments. Only 25.1% of them can be explained by word for word correspondences. It means that we need a statistical model which can handle phrasal mappings.

In the case of the phrasal mappings, a lot of parameters should be searched even if we restrict the length of word strings. Moreover, in order to properly estimate parameters we need much larger volume of bilingual aligned text than it in word-for-word modeling. Even though such a large corpora exist sometimes, they do not come up with the lexical alignments.

For this problem, we here consider syntactic features which are important in determining structures. A structural feature means here a mapping between tag sequences in bilingual parallel sentences.

If we are concerned with tag sequence alignments, it is possible to estimate statistical parameters in a relatively small size of corpora. As a result, we can remarkably reduce the problem space for possible lexical alignments, a sort of t probability in (1), which improve the complexity of a statistical machine translation model.

If there are similarities between corresponding tag sequences in two language, the structural features would be easily computed or recognized. However, a tag sequence in English can be often translated into a completely different tag sequence in Korean as follows.

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

It means that similarities of tag features between two languages are not kept all the time and it is necessary to get the most likely tag sequence mappings that reflect structural correspondences between two languages.

In this paper, the tag sequence mappings are obtaind by automatic feature selection based on the maximum entropy model.

3 Problem Setting

In this chapter, we describe how the features are related to the training data. Let t_e be an English tag sequence and t_k be a Korean tag sequence. Let $\mathcal{T}_S$ be the set of all possible tag sequence mappings in a aligned sentence, S. We define a feature function (or a feature) as follows:

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

It indicates co-occurrence information between tags appeared in $\mathcal{T}_S$. $f(t_e, t_k)$ expresses the information for predicting that t_e maps into t_k. A feature means a sort of information for predicting something. In our model, co-occurrence information on the same aligned sentence is used for a feature, while context is used as a feature in most of systems using maximum entropy. It can be less informative than context. Hence, we considered an initial supervision and feature selection.

Our model starts with initial seed(active) features for mapping extracted by supervision. In the next step, feature pool is constructed from training samples from filtering and only features with a large gain to the model are added into active feature set. The final outputs of our model are the set of active features, their gain values, and conditional probabilities of features which maximize the model. The results can be embedded in parameters of statistical machine translation and help to construct structural bilingual text.

Most alignment algorithm consists of two steps:
(1) estimate translation probabilities.
(2) use these probabilities to search for most probable alignment path.
Our study is focused on (1), especially the part of tag string alignments.

Next, we will explain the concept of the model. We are concerned with an optimal statistical model which can generate the training samples. Namely, our task is to construct a stochastic model that pro-

duces output tag sequence $\mathcal{T}_k$, given a tag sequence $\mathcal{T}_e$. The problem of interest is to use samples of tagged sentences to observe the behavior of the random process. The model p estimates the conditional probability that the process will output t_e, given t_k. It is chosen out of a set of all allowed probability distributions.

The following steps are employed for our model.

Input: a set **L** of POS-labeled bilingual aligned sentences.

1. Make a set $\mathcal{F}$ of correspondence pairs of tag sequences, (t_e, t_k) from a small portion of **L** by supervision.

2. Set $\mathcal{F}$ into a set of active features, $\mathcal{A}$.

3. Maximization of parameters, λ of active features by IIS(Improved Iterative Scaling) algorithm.

4. Create a feature pool set $\mathcal{P}$ of all possible alignments $\mathbf{a}(t_e, t_k)$ from tag sequences of samples.

5. Filter $\mathcal{P}$ using frequency and similarity with $\mathcal{A}$.

6. Compute the approximate gains of features in $\mathcal{P}$.

7. Select new features($\mathcal{N}$) with a large gain value, and add $\mathcal{A}$.

Output: $p(t_k|t_e) where (t_e, t_k) \in \mathcal{A}$ and their λ_i.

We began with training samples composed of English-Korean aligned sentence pairs, $(\mathbf{e}, \mathbf{k})$. Since they included long sentences, we broke them into shorter ones. The length of training sentences was limited to under 14 on the basis of English. It is reasonable because we are interested in not lexical alignments but tag sequence alignments. The samples were tagged using brill's tagger and 'Morany' that we implemented as a Korean tagger. Figure 1 shows the POS tags we considered. For simplicity, we adjusted some part of Brill's tag set.

In the supervision step, 700 aligned sentences were used to construct the tag sequences mappings which are referred to as an active feature set $\mathcal{A}$. As Figure 2 shows, there are several ways in constructing the correspondences. We chose the third mapping although (1) can be more useful to explain Korean with predicate-argument structure. Since a subject of a English sentence is always used for a subject form in Korean, we exlcuded a subject case from arguments of a predicate. For example, 'they' is only used for a subject form, whereas 'me' is used for a object form and a dative form.

In the next step, training events, (t_e, t_k) are constructed to make a feature pool from training samples. The event consists of a tag string t_e of a English
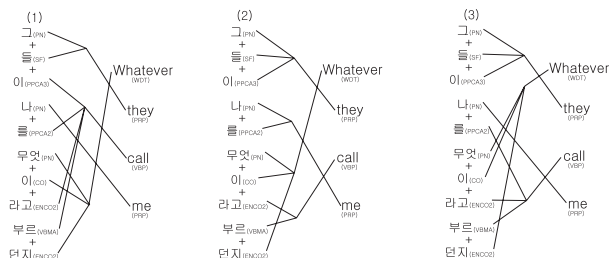

Figure 2: Tag sequence correspondences at the phrase level

POS-tagged sentence and a tag string t_k of the corresponding Korean POS-tagged sentence and it can be represented with indicator functions $f_i(t_e, t_k)$.

For a given sequence, the features were drawn from all adjacent possible pairs and some interrupted pairs. Only features $(t_{e\,i}, t_{k\,i})$ out of the feature pool that meet the following conditions are extracted.

- $\#(t_{e\,i}, t_{k\,i}) \geq 3$, $\#$ is count
- there exist $t_{k\,x}$, where $(t_{e\,i}, t_{k\,x})$ in $\mathcal{A}$ and the similarity(same tag count) of $t_{k\,i}$ and $t_{k\,x} \geq 0.6$

Table 1 shows possible features, for a given aligned sentence , 'take her out − *gnyeoreul baggeuro deryeogara*'.

Since the set of the structural features for alignment modeling is vast, we constructed a maximum entropy model for $p(t_k|t_e)$ by the iterative model growing method.

4 Maximum Entropy

To explain our method, we briefly describe the concept of maximum entropy. Recently, many approaches based on the maximum entropy model have been applied to natural language processing (Berger et al., 1994; Berger et al., 1996; Pietra et al., 1997).

Suppose a model **p** which assigns a probability to a random variable. If we don't have any knowledge, a reasonable solution for **p** is the most uniform distribution. As some knowledge to estimate the model **p** are added, the solution space of **p** are more constrained and the model would be close to the optimal probability model.

For the purpose of getting the optimal probability model, we need to maximize the uniformity under some constraints we have. Here, the constraints are related with features. A feature, f_i is usually represented with a binary indicator function. The importance of a feature, f_i can be identified by requiring that the model accords with it.

As a constraint, the expected value of f_i with respect to the model $p(f_i)$ is supposed to be the same as the expected value of f_i with respect to empirical distribution in training samples, $\tilde{p}(f_i)$.

TAG	DESCRIPTION	TAG	DESCRIPTION
.	comma	.	sentence terminator
CC	conjunction,coordinating	CD	numeral, cardinal
DT	determiner	EX	existential there
FW	foreign word	IN	preposition, subordinating
JJ	adjective, ordinal	JJR	adjective, comparative
JJS	adjective, superlative	LS	list item marker
MD	modal auxiliary	NN	noun, common
NNP	noun, proper, singular	NNPS	noun, proper, plural
PDT	pre-determiner	POS	genitive marker
PRP	pronoun, personal	PRP$	pronoun, possessive
RB	adverb	RBR	adverb, comparative
RBS	adverb, superlative	RP	particle
SYM	symbol	TO	to or infinitive marker
UH	interjection	VBP	verb, present tense
VBD	verb, past tense	VBG	verb, present participle
VBN	verb, past participle	WDT	WH-determiner
WP$	WH-pronoun, possessive	WRB	WH-adverb
NOT	not	BEP	be verb, present tense
BED	be verb, past tense	BEN	be verb, past participle
BEG	be verb, present participle	HVP	have verb, present tense
HVD	have verb, past participle	DOP	do verb, present tense
DOD	do verb, past tense	DON	do verb, past participle

TAG	POS	TAG	POS
NNIN1	proper noun	PPCA1	nominative postposition
NNIN2	common noun	PPCA2	accusative postposition
NNDE1	common-dependent noun	PPCA3	possessive postposition
NNDE2	unit-dependent noun	PPCA4	vocative postposition
PN	pronoun	PPAD	adverbial postposition
NU	number	PPCJ	conjunctive postposition
VBMA	verb	PPAU	auxiliary postposition
AJMA	adjective	ENTE	final ending
CO	copula	ENCO1	coordinate ending
AX	auxiliary verb	ENCO2	subordinate ending
ADCO	constituent adverb	ENCO3	auxiliary ending
ADSE	sentential adverb	ENTR1	adnominal ending
CJ	conjunctive adverb	ENTR2	nominal ending
ANCO	configurative adnominal	ENTR3	adverbial ending
ANDE	demonstrative adnominal	ENCM	ending+postposition
ANNU	numeral adnominal	PE	pre-ending
EX	exclamation	SF	suffix
LQ	left quotation mark	PF	prefix
RQ	right quotation mark	CM	comma
SY	symbols	SC	termination

Figure 1: English Tags (left) and Korean Tags (right)

English Tag Sequences	Korean Tag Sequences
[VBP+IN] [take+out] [1+3]	[PPCA2+PPAD+VBMA] $[reul+euro+deryeoga]$ [2+4+5]
[VBP] [take] [1]	[PN] $[geunyeo]$ [1]
[VBP+PRP] [take+her] [1+2]	[PPAD+VBMA+ENTE] $[reul+euro+deryeoga+ra]$ [4+5+6]
[VBP+PRP+IN] [take+her+out] [1+2+3]	[NNIN2] $[bagg]$ [3]
[PRP] [her] [2]	[NNIN2+PPAD] $[bagg+euro]$ [3+4]
[IN] [out] [3]	[ENTE] $[ra]$ [6]
	[PPAD+VBMA] $[euro+deryeoga]$ [4+5]
	[PPAD+VBMA+ENTE] $[euro+deryeoga+ra]$ [4+5+6]
	[PPCA2+NNIN2+PPAD+VBMA] $[reul+bagg+euro+deryeoga]$ [2+3+4+5]
	[PPCA2+NNIN2+PPAD+VBMA+ENTE] $[reul+bagg+euro+deryeoga+ra]$ [2+3+4+5+6]
	[PPCA2+NNIN2+PPAD+VBMA] $[reul+deryeoga]$ [2+3+4+5]
	[PPCA2+NNIN2+PPAD+VBMA+ENTE] $[reul+deryeoga+ra]$ [2+3+4+5+6]

Table 1: possible tag sequences

In sum, the maximum entropy framework finds the model which has highest entropy(most uniform), given constraints. It is related to the constrained optimization. To select a model from a constrained set C of allowed probability distributions, the model $p_\star \in C$ with maximum entropy $H(p)$ is chosen.

In general, for the constrained optimization problem, Lagrange multipliers of the number of features can be used. However, it was proved that the model with maximum entropy is equivalent to the model that maximizes the log likelihood of the training samples like (2) if we can assume it as an exponential model.

In (2), the left side is Lagrangian of the conditional entropy and the right side is maximum log-likelihood. We use the right side equation of (2) to select $\lambda_\star$ for the best model $p_\star$.

$$argmax_{\lambda_i}(-\sum_{x,y}\bar{p}(x)p(y|x)log p(y|x)+\lambda_i(p(f_i)-\bar{p}(f_i))) \quad (2)$$
$$=argmax_{\lambda_i}\sum_{x,y}\bar{p}(x,y)log p(y|x)$$

Since $\lambda_\star$ cannot be found analytically, we use the following improved iterative scaling algorithm to compute $\lambda_\star$ of n active features in $\mathcal{A}$ in total samples.

1. Start with $\lambda_i = 0$ for all $i \in \{1, 2, \ldots, n\}$

2. Do for each $i \in \{1, 2, \ldots, n\}$:

   (a) Let $\Delta\lambda_i$ be the solution to the log likelihood

   (b) Update the value of λ_i into $\lambda_i + \Delta\lambda_i$, where $\Delta\lambda_i = \log \frac{\sum_{x,y}\bar{p}(x,y)f_i(x,y)}{\sum_{x,y}\bar{p}(x)p_\lambda(y|x)f_i(x,y)}$
   $p_\lambda(y|x) = \frac{1}{Z_\lambda(x)}e^{(\sum_i \lambda_i f_i(x,y))}$,
   $Z_\lambda(x) = \sum_y e^{(\sum_i \lambda_i f_i(x,y))}$

3. Stop if not all the λ_i have converged, otherwise go to step 2

The exponential model is represented as (3). Here, λ_i is the weight of feature f_i. In our model, since only one feature is applied to each pair of x and y, it can be represented as (4) and f_i is the feature related with x and y.

$$\tilde{p}(y|x) = \frac{\sum_i e^{\lambda_i f_i(x,y)}}{\sum_y e^{\sum_i \lambda_i f_i(x,y)}} \quad (3)$$

$$\tilde{p}(y|x) = \frac{e^{\lambda_i f_i(x,y)}}{\sum_y e^{\lambda_i f_i(x,y)}} \quad (4)$$

5 Feature selection

Only a small subset of features will be employed in a model by selecting useful features from the feature pool $\mathcal{P}$. Let $p_{\mathcal{A}}$ be the optimal model constrained by a set of active features $\mathcal{A}$ and $\mathcal{A} \cup f_i$ be $\mathcal{A}f_i$. Let $p_{\mathcal{A}f_i}$ be the optimal model in the space of probability distribution $C(\mathcal{A}f_i)$. The optimal model can be represented as (5). Here, the optimal model means a maximum entropy model.

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

$$Z_{\alpha}(x) = \sum_y p_{\mathcal{A}}(y|x) e^{\alpha f_i(x,y)} \quad (5)$$

The improvement of the model regarding the addition of a single feature f_i can be estimated by measuring the difference of maximum log-likelihood between $L(p_{\mathcal{A}f_i})$ and $L(p_{\mathcal{A}})$. We denote the gain of feature f_i by $\Delta(\mathcal{A}f_i)$ and it can be represented in (6).

$$
\begin{aligned}
\Delta(\mathcal{A}f_i) &\equiv max_{\alpha} G_{\mathcal{A}f_i}(\alpha) \\
G_{\mathcal{A}f_i}(\alpha) &\equiv L(p_{\mathcal{A}f_i}) - L(p_{\mathcal{A}}) \\
&= -\sum_x \tilde{p}(x) \sum_y p_{\mathcal{A}}(y|x) e^{\alpha f_i(x,y)} \\
&\quad + \alpha \tilde{p}(f_i) \quad (6)
\end{aligned}
$$

Note that a model $p_{\mathcal{A}}$ has a set of parameters λ which means weights of features. The model $p_{\mathcal{A}f_i}$ contains the parameters and the new parameter α with respect to the feature f_i. When adding a new feature to $\mathcal{A}$, the optimal values of all parameters of probability distribution change. To make the computation of feature selection tractable, we approximate that the addition of a feature f_i affects only the single parameter α, as shown in (5).

The following algorithm is used for computing the gain of the model with respect to f_i. We referred to the studies of (Berger et al., 1996; Pietra et al., 1997). We skip the detailed contents and proofs.

---

1. Let
$$r = \begin{cases} 1 & if \ \tilde{p}(f_i) \leq p_{\mathcal{A}}(f_i) \\ -1 & otherwise \end{cases}$$

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
$$G_{\mathcal{A}f_i}(\alpha) = -\sum_x \tilde{p}(x) \log Z_{\alpha}(x) + \alpha \tilde{p}(f_i) \ ,$$
$$G'_{\mathcal{A}f_i}(\alpha) = \tilde{p}(f_i) - \sum_x \tilde{p}(x) M(x) \ ,$$
$$G''_{\mathcal{A}f_i}(\alpha) = -\sum_x \tilde{p}(x) p^{\alpha}_{\mathcal{A}f_i}((f_i - M(x))^2|x)$$

set	description	# of disjoint features	total events
A	active features	1483	4113
P	feature candidates	3172	63773
N	new features	97	5503

Table 2: Summery of Features Selected

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. We can recognize the importance of a feature with the gain value. As mentioned above, it means how much the feature accords with the model. We viewed the feature as the information that t_k and t_e occur together.

6 Experimental results

The total samples consists of 3,000 aligned sentence pairs of English-Korean, which were extracted from news on the web site of 'Korea Times' and a magazine for English learning.

In the initial step, we manually constructed the correspondences of tag sequences with 700 POS-tagged sentence pairs. In the supervision step, we extracted 1,483 correct tag sequence correspondences as shown in Table 2, and it work as active features. As a feature pool, 3,172 disjoint features of tag sequence mappings were retrieved. It is very important to make atomic features.

We maximized λ of active features with respect to total samples using improved the iterative scaling algorithm. Figure 3 shows λ_i of each feature $f(t_{BEP+JJ}, t_k) \in \mathcal{A}$. There are many correspondence patterns with respect to the Englsh tag string, 'BEP+JJ'.

Note that $p(t_k|t_e)$ is computed by the exponential model of (4) and the conditional probability is the same with empirical probability in (7). Since the value of $p(y|x)$ shows the maximum likelihood, it is proved that each λ was converged correctly.

$$p(y|x) \equiv \frac{\# \ of \ (x,y) \ occurs \ in \ sample}{number \ of \ times \ of \ x} \quad (7)$$

In feature selection step, we chose useful features with the gain threshold of 0.008. Figure 4 shows some feaures with a large gain. Among them, tag sequences mapping including 'RB' are erroneous. It means that position of adverb in Korean is very complicated to handle. Also, proper noun in English aligned common nouns in Korean

Feature(x,y)		λ	c(x,y)	p(y|x)	Example	
English	Korean				English	Korean
BEP+JJ	VBMA+ENCO3+AX+ENTE	10.1369	162.00	0.4247	are+prepared	준비되+어+있+ㅂ니다
BEP+JJ	VBMA	8.8520	45.00	0.1180	are+careful	주의하
BEP+JJ	AJMA	8.6787	39.00	0.0996	am+healthy	건강하
BEP+JJ	AJMA+ENTE	8.2628	25.00	0.0655	is+new	새롭+다
BEP+JJ	VBMA+ENTE	7.2379	9.00	0.0236	am+sure	확신하+ㅂ니다
BEP+JJ	NNIN2+CO	7.1372	8.00	0.0210	am+rich	부자+이
BEP+JJ	NNIN2+CO+VBMA	6.9909	7.00	0.0183	is+selfish	이기적+이+ㅂ니다
BEP+JJ	NNIN2+PPCA1+VBMA+ENTE	6.8402	6.00	0.0157	is+patriotic	애국자+가+되+다
BEP+JJ	NNIN2+CO+ENTE	6.8308	6.00	0.0157	is+reasonable	합리적+이+다
BEP+JJ	NNIN2+PPCA2+AX+ENTE	6.4256	4.00	0.0105	is+reprehensible	비난받+을+만하+ㅂ니다
BEP+JJ	NNIN2+PPCA1+VBMA	6.4250	4.00	0.0105	is+helpful	도움+이+되

Figure 3: λ of active features in 𝒜

Figure 5: Best Lexical alignment

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 4: Conditional Probability
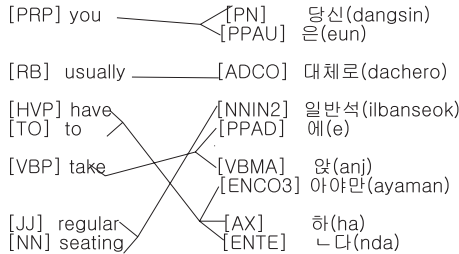
because of tagging errors. Note that in the case of 'PN+PPCA2+PPAD+VBMA', it is not an adjacent string but an interrupted string. It means that a verb in English generally map to a verb taking as argument the accusative and adverbial postposition in Korean.

One way of testing usefulness of our method is to construct structured aligned bilingual sentences. Table 3 shows lexical alignments using tag sequence alignments drawn from our algorithm for a given sentence, 'you usually have to take regular seating - *dangsineun dachero ilbanseoke anjayaman handa*' and Figure 5 shows the best lexical alignment of the sentence.

We conducted the experiment on 100 sentences composed of words in length 14 or less and simply chose the most likely paths. As the result, the accuray was about 71.1%. It shows that we can partly use the tag sequence alignments for lexical alignments. We will extend the structural mapping model with consideration to the lexical information. The parameters, the conditional probabilities about stuctural mappings will be embedded in a statistical model. Table 4 shows conditional probabilities of some features according to 'DT+NN'. In general, determiner is translated into NULL or adnominal word in Korean.

7 Conclusion

When aligning English-Korean sentences, the differences of word order and word unit require structural information. For this reason, we tried structural tag string mapping using maximum entropy modeling and feature selection concept. We devised a model that generates a English tag string given a Korean tag string. From initial active structural features, useful features are extended by feature selection. The retrieved features and parameters can be embedded in statistical machine translation and reduce the complexity of searching. We showed that they can helpful to construct structured aligned bilingual sentences.

References

Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillett, John D. Lafferty, Robert L. Mercer, Harry Printz, and Lubos Ures. 1994. The Candie system for machine translation. In *Proceedings of the ARPA Conference on Human Language Technology*, Plainsborough, New Jersey.

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.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79-85

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer. 1993. The mathematics of statistical machine translation: pa-

Feature(x,y) x	Y	σ̂	P(y|x)	△L(Afi)	Example English	Korean
VBP+PRP+TO	PN+PPCA2+PPAD+VBMA	9.8687	0.1722	0.0194	send+him+to	그+를+~에+보내
BEP+RBR+IN	PPAD+AJMA	9.6780	0.3265	0.0192	is+more+than	~보다+많
DT+CD	NU+NNDE2	9.2799	0.2449	0.0190	the+two	두+명
JJ+IN	PPAD+AJMA+ENTR1	9.5343	0.2450	0.0190	smarter+than	보다+영리+하+ㄴ
VBG+TO	PPAD+PPCA2+VBMA	9.9542	0.3269	0.0189	serving+to	~에게+~을+나르
BEP	PPCA1+AJMA	9.6720	0.2941	0.0188	is	~이+있
BEP	PPCA1+AJMA+ENTE	9.2724	0.1961	0.0188	are	~이+있+다
NNP	NNIN2+PPAU	8.7481	0.1225	0.0182	IBM	IBM+은
NNP	NNIN2+NNIN2	9.1397	0.1337	0.0180	Harvard	하버드+대학
TO+PRP	PN+PPAD	9.5634	0.2307	0.0180	to+him	그+에게
TO+PRP	PN+PPCA1	9.2604	0.1730	0.0180	to+her	그녀+가
MD+RB	ENTR1+NNDE1+CO	9.2445	0.1548	0.0177	?	?
MD+RB	ENTR1+NNDE1+CO+ENTE	9.2564	0.1548	0.0177	?	?
MD+BEP	CO+ENCO2+VBMA	8.5435	0.0934	0.0177	should+be	이+어야+하
NNP+NNS	NNIN2+SF	9.1597	0.1470	0.0176	English+books	영어책+들
VBP+TO+VBP	PPCA2+VBMA+ENCO2+VBMA	8.9928	0.1278	0.0174	request+to+send	~를+보내+라고+요청하
BED+VBN+IN	PPAD+VBMA+PE+ENTE	9.1511	0.1704	0.0173	was+thrown+to	~에게+던져지+었+다
BED+VBN+IN	PPAD+VBMA+PE	9.1636	0.1705	0.0173	were+sent+to	

Figure 4: Some features with a large gain

Tag alignment	Conditional	Lexical alignment
PRP : PN+PPAU	0.150109	you : dangsin+eun
RB : ADCO	0.142193	usually : dachero
RB : NNIN2+PPAD	0.038105	usually : ilbanseok+e
HVP+TO : ENCO3+AX+ENTE	0.982839	have+to : ayaman+handa
VBP : PPAD+VBMA	0.050224	take : e+anj
VBP : VBMA+ENCO3+AX+ENTE	0.011110	take : anjay+aman+ha+nda
VBP : PPAD+VBMA+ENCO3+AX+ENTE	0.001851	take : e+anjayaman+handa
VBP+JJ : NNIN2+PPAD+VBMA	0.057657	take+regular : ilbanseok+e+anj
JJ+NN : NNIN2	0.581791	regular+seating : ilbanseok

Table 3: Lexical alignments using tag alignments

rameter estimation. *Computational Linguistics*, 19(2):263-311.

Stanley F. Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of ACL 31*, 9-16.

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.

William A. Gale, 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.

Marin Kay, Martin Roscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19:121-142.

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

Yuji Matsumoto, Hiroyuki Ishimoto, Takehito Utsuro. 1993. Structural matching of parallel texts. In *Proceedings of ACL 31*, 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 Hans Weber. 1998. Improving Statistical Natural Language Translation with Categories and Rules. In *Proceedings of ACL 36/COLING*, 985-989.

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.

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

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*.

C. Tillmann, S. Vogel, H. Ney, and A. Zubiaga. 1997. A DP based search using monotone alignments in statistical translation. In *Proceedings of ACL 35/EACL 8*, 289-296.

Ye-Yi Wang and Alex Waibel. 1997. Decoding algorithm in statistical machine translation. In *Proceedings of ACL 35/EACL 8*, 366-372.

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

Dekai Wu 1996. A polynomial-time algorithm for statistical machine translation. In *Proceeding of ACL 34*.