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 languages, 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 give 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 entropy 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)

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

(1)

Here, \( n \) is the fertility probability that an English word generates \( n \) French words, \( t \) is the 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. This formula is one of many ways in which \( p(f, a|c) \) 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-forward 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.

\[ \text{can/MD} \rightarrow \sim \text{eul/ENTR1} \text{su/NNDE1} \text{'iss/AJMA da/ENTE} \]

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 obtained 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 \( T_S \) be the set of all possible tag sequence mappings in an aligned sentence \( S \). We define a feature function (or a feature) as follows:

\[ f(t_e, t_k) = \begin{cases} 1 & \text{pair}(t_e, t_k) \in T_S \\ 0 & \text{otherwise} \end{cases} \]

It indicates co-occurrence information between tags appeared in \( T_S \). \( f(t_e, t_k) \) 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 $T_e$, given a tag sequence $T_c$. 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 $T_c$. 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 $F$ of correspondence pairs of tag sequences, $(t_e, t_k)$ from a small portion of $L$ by supervision.

2. Set $F$ into a set of active features, $A$.

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

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

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

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

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

**Output**: $p(t_k|t_e)$ where $(t_e, t_k) \in A$ and their $\lambda_i$.

We began with training samples composed of English-Korean aligned sentence pairs, (e,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 $A$. 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 excluded 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 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 sentence, the features were drawn from all adjacent possible pairs and some interrupted pairs. Only features $(t_e, t_k)$ out of the feature pool that meet the following conditions are extracted.

- $\#(t_e, t_k) \geq 3$, $\#$ is count
- there exist $t_kx$, where $(t_e, t_kx) \in A$ and the similarity (same tag count) of $(t_kx, t_kx) \geq 0.6$

Table 1 shows possible features for a given aligned sentence, 'take her out - gyeoreul baggyeol deryeongara'.

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, $\bar{p}(f_i)$.
In summary, 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 of allowed probability distributions, the model \( p_* \in C \) 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 \( \lambda_1 \) for the best model \( p_* \).

\[
\arg \max \sum_{x,y} \left\{ -\sum_{x,y} p(x,y) \log p(x,y) + \lambda_1 (f_1(x,y) - \hat{p}(f_1)) \right\} \tag{2}
\]

Since \( \lambda_i \) cannot be found analytically, we use the following improved iterative scaling algorithm to compute \( \lambda_* \) of a active features in \( A \) 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 \( \lambda_i \) into \( \lambda_i + \Delta \lambda_i \),

   where \( \Delta \lambda_i = \log \sum_{x,y} p(x,y) f_i(x,y) \),

   \[
p_i(y|x) = \frac{1}{Z_i(x)} e^{\lambda_i f_i(x,y)}
   \]

   \[
   Z_i(x) = \sum_{y} e^{\lambda_i f_i(x,y)}
   \]

3. Stop if not all the \( \lambda_i \) have converged, otherwise go to step 2.
5 Feature selection

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

\[
\begin{align*}
  p_{A_{f_{i}}} & = \frac{1}{Z_{\alpha}(x)} p_{A}(y|x) e^{\alpha f_{i}(x,y)} \\
  Z_{\alpha}(x) & = \sum_{y} p_{A}(y|x) e^{\alpha f_{i}(x,y)}
\end{align*}
\]

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 \( L(p_{A_{f_{i}}}) \) and \( L(p_{A}) \). We denote the gain of feature \( f_{i} \) by \( \Delta(A_{f_{i}}) \) and it can be represented in \( (5) \).

\[
\begin{align*}
  \Delta(A_{f_{i}}) & \equiv \max_{\alpha} G_{A_{f_{i}}}(\alpha) \\
  G_{A_{f_{i}}}(\alpha) & \equiv I(p_{A_{f_{i}}}) - I(p_{A}) \\
  & = -\sum_{x} \bar{p}(x) \sum_{y} p_{A}(y|x) e^{\alpha f_{i}(x,y)} \\
  & \quad + \alpha \bar{p}(f_{i})
\end{align*}
\]

Note that a model \( p_{A} \) has a set of parameters \( \lambda \) which means weights of features. The model \( p_{A_{f_{i}}} \) contains the parameters and the new parameter \( \alpha \) with respect to the feature \( f_{i} \). When adding a new feature to \( A \), 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 \( \alpha_{i} \), 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 \( r = \left\{ \begin{array}{ll} 1 & \text{if } \bar{p}(f_{i}) \leq p_{A}(f_{i}) \\ -1 & \text{otherwise} \end{array} \right. \)
2. Set \( \alpha_{0} = 0 \)
3. Repeat the following until \( G_{A_{f_{i}}}(\alpha_{n}) \) has converged:
   1. Compute \( \alpha_{n+1} \) from \( \alpha_{n} \) using
      \[
      \alpha_{n+1} = \alpha_{n} + \frac{1}{r} \log \left( 1 - \frac{1}{r} G_{A_{f_{i}}}(\alpha_{n}) \right)
      \]
   2. Compute \( G_{A_{f_{i}}}(\alpha_{n+1}) \) using
      \[
      \begin{align*}
      G_{A_{f_{i}}}(\alpha) & = -\sum_{x} \bar{p}(x) \log Z_{\alpha}(x) + \alpha \bar{p}(f_{i}) \\
      G_{A_{f_{i}}}'(\alpha) & = \bar{p}(f_{i}) - \sum_{x} \bar{p}(x) M(x) \\
      G_{A_{f_{i}}}''(\alpha) & = -\sum_{x} \bar{p}(x) p_{A_{f_{i}}}'((f_{i} - M(x)))^{2} |x|
      \end{align*}
      \]

4. Set \( \Delta L(A_{f_{i}}) \leftarrow G_{A_{f_{i}}}(\alpha_{n}) \)

This algorithm is iteratively computed using Newton’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 \( \alpha \) and \( t_{k} \) 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 \( \lambda \) of active features with respect to total samples using improved the iterative scaling algorithm. Figure 3 shows \( \lambda \) of each feature \( f(k_{BRP+JJ},t_{k}) \in A \). There are many correspondence patterns with respect to the English tag string, ‘BRP+JJ’.

Note that \( p(t_{k}|x) \) 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 \( \lambda \) was converged correctly.

\[
p(y|x) = \frac{\# \text{ of } (x,y) \text{ occurs in sample}}{\text{number of times of } x}
\]

In feature selection step, we chose useful features with the gain threshold of 0.008. Figure 4 shows some features 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.
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 maps 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’ 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 accuracy 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 structural 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.

### References

Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillet, John D. Lafferty, Robert L. Mercer, Harry Printz, and Lubes Ures. 1994. The Candide 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.
Figure 4: Some features with a large gain

Table 3: Lexical alignments using tag alignments

Parameter 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 Ishinoto, 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 Hatziwassiloglou. 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*.

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

445