Discriminative Induction of Sub-Tree Alignment using Limited Labeled Data

Jun Sun[1,2] **Min Zhang**[1] **Chew Lim Tan**[2]
[1]Institute for Infocomm Research [2]School of Computing, National University of Singapore
sunjun@comp.nus.edu.sg mzhang@i2r.a-star.edu.sg tancl@comp.nus.edu.sg

Abstract

We employ Maximum Entropy model to conduct sub-tree alignment between bilingual phrasal structure trees. Various lexical and structural knowledge is explored to measure the syntactic similarity across Chinese-English bilingual tree pairs. In the experiment, we evaluate the sub-tree alignment using both gold standard tree bank and the automatically parsed corpus with manually annotated sub-tree alignment. Compared with a heuristic similarity based method, the proposed method significantly improves the performance with only limited sub-tree aligned data. To examine its effectiveness for multilingual applications, we further attempt different approaches to apply the sub-tree alignment in both phrase and syntax based SMT systems. We then compare the performance with that of the widely used word alignment. Experimental results on benchmark data show that sub-tree alignment benefits both systems by relaxing the constraint of the word alignment.

1 Introduction

Recent research in Statistical Machine Translation (SMT) tends to incorporate more linguistically grammatical information into the translation model known as linguistically motivated syntax-based models. To develop such models, the phrasal structure parse tree is usually adopted as the representation of bilingual sentence pairs either on the source side (Huang et al., 2006; Liu et al., 2006) or on the target side (Galley et al., 2006; Marcu et al., 2006), or even on both sides (Graehl and Knight, 2004; Zhang et al., 2007). Most of the above models either construct a pipeline to transform from/to tree structure, or synchronously generate two trees in parallel (i.e., synchronous parsing). Both cases require syntactically rich translational equivalences to handle non-local reordering. However, most current works obtain the syntactic translational equivalences by initially conducting alignment on the word level. To employ word alignment as a hard constraint for rule extraction has difficulty in capturing such non-local phenomena and will fully propagate the word alignment error to the later stage of rule extraction.

Alternatively, some initial attempts have been made to directly conduct syntactic structure alignment. As mentioned in Tinsley et al. (2007), the early work usually constructs the structure alignment by hand, which is time-consuming. Recent research tries to automatically align the bilingual syntactic sub-trees. However, most of these works suffer from the following problems. Firstly, the alignment is conducted based on heuristic rules, which may lose extensibility and generality in spite of accommodating some common cases (Groves et al., 2004). Secondly, various similarity computation methods are used based merely on lexical translation probabilities (Tinsley et al., 2007; Imamura, 2001) regardless of structural features. We believe the structure information is an important issue to capture the non-local structural divergence of languages by modeling beyond the plain text.

To address the above issues, we present a statistical framework based on Maximum Entropy (MaxEnt) model. Specifically, we consider sub-tree alignment as a binary classification problem and use Maximum Entropy model to classify each instance as *aligned* or *unaligned*. Then, we perform a greedy search within the reduced search space to conduct sub-tree alignment links based on the alignment probabilities obtained from the classifier.

Unlike the previous approaches that can only measure the structural divergence via lexical features, our approach can incorporate both lexical and structural features. Additionally, instead of explicitly describing the instances of sub-tree pairs as factorized sub-structures, we frame most of our features as score based feature functions, which helps solve the problem using limited sub-tree alignment annotated data. To train the model and evaluate the alignment performance, we adopt

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1047–1055,
Beijing, August 2010

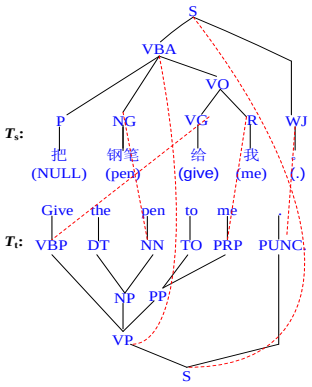

Figure 1: Sub-tree alignment as referred to Node alignment

HIT Chinese-English parallel tree bank for gold standard evaluation. To explore its effectiveness in SMT systems, we also manually annotate sub-tree alignment on automatically parsed tree pairs and perform the noisy data evaluation. Experimental results show that by only using limited sub-tree aligned data of both corpora, the proposed approach significantly outperforms the baseline method (Tinsley et al., 2007). The proposed features are very effective in modeling the bilingual structural similarity. We further apply the sub-tree alignment to relax the constraint of word alignment for both phrase and syntax based SMT systems and gain an improvement in BLEU.

2 Problem definition

A sub-tree alignment process pairs up the sub-trees across bilingual parse trees, whose lexical leaf nodes covered are translational equivalent, i.e., sharing the same semantics. Grammatically, the task conducts links between syntactic constituents with the maximum tree structures generated over their word sequences in bilingual tree pairs.

In general, sub-tree alignment can also be interpreted as conducting multiple links across internal nodes between sentence-aligned tree pairs as shown in Fig. 1. The aligned sub-tree pairs usually maintain a non-isomorphic relation with each other especially for higher layers. We adapt the same criteria as Tinsley et al. (2007) in our study:

(i) a node can only be linked once;
(ii) descendants of a source linked node may only link to descendants of its target linked counterpart;
(iii) ancestors of a source linked node may only link to ancestors of its target linked counterpart.

where the term "node" refers to root of a sub-tree, which can be used to represent the sub-tree.

3 Model

We solve the problem as binary classification and employ MaxEnt model with a greedy search.

Given a bilingual tree pair S^I and T^J, $S^I = \{s_1, \dots, s_i, \dots, s_I\}$ is the source tree consisting of I sub-trees, where I is also the number of nodes in the source tree S^I. $T^J = \{t_1, \dots, t_j, \dots, t_J\}$ is the target tree consisting of J sub-trees, where J is also the number of nodes in the target tree T^J.

For each sub-tree pair (s_i, t_j) in the given bilingual parse trees (S^I, T^J), the sub-tree alignment probability is given by:

$$Pr(a|s_i, t_j, \theta) = \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(a, s_i, t_j, \theta)]}{\sum_{a'} \exp[\sum_{m=1}^{M} \lambda_m h_m(a', s_i, t_j, \theta)]} \quad (1)$$

where

$$a = \begin{cases} 1 & \text{if } (s_i, t_j) \text{ is aligned} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Feature functions are defined in a quadruple (a, s_i, t_j, θ). θ is an additional variable to incorporate new dependencies other than the sub-tree pairs. For each feature function $h_m(a, s_i, t_j, \theta)$, a weight λ_m is applied to tailor the distribution.

After classifying the candidate sub-tree pairs as *aligned* or *unaligned*, we perform a greedy search within the reduced search space to conduct *sure* links based on the conditional probability $Pr(a|s_i, t_j, \theta)$ obtained from the classifier. The alignment probability is independently normalized for each sub-tree pair and hence suitable as a searching metric.

The greedy search algorithm can be described as an automaton. A state in the search space is a partial alignment with respect to the given bilingual tree pair. A transition is to add one more link of node pairs to the current state. The initial state has no link. The terminal state is a state where no more links can be added according to the definition in Section 2. We use greedy search to generate the best-links at the early stage. There are cases that the correctly-aligned tree pairs have very few links, while we have a bunch of candidates with lower alignment probabilities. However, the sum of the lower probabilities is larger than that of the correct links', since the number of correct links is much fewer. This makes the alignment results biased to be with more links. The greedy search helps avoid this asymmetric problem.

4 Feature Functions

In this section, we introduce a variety of feature functions to capture the semantically equivalent

counterparts and structural divergence across languages. For the semantic equivalence, we define lexical and word alignment feature functions. Since those feature functions are directional, we describe most of these functions as conditional feature functions based on the conditional lexical probabilities. We also introduce the tree structural features to deal with the structural divergence of bilingual parse trees. Inspired by Burkett and Klein (2008), we introduce the feature functions in an internal-external manner based on the fact that the feature scores for an aligned sub-tree pair tend to be high inside both sub-trees, while they tend to be low inside one sub-tree and outside the other.

4.1 Internal Lexical Features

We use this feature to measure the degree of semantic equivalence of the sub-tree pair. According to the definition of sub-tree alignment in Section 2, the word sequence covered by the sub-tree pair should be translational equivalence. Therefore, the lexicons within the two corresponding sub-spans should be highly related in semantics. We define the internal lexical features as follows:

$$\phi(s_i|t_j) = \left(\prod_{v\in in(t_j)} \sum_{u\in in(s_i)} P(u|v)\right)^{\frac{1}{|in(t_j)|}}$$

$$\phi(t_j|s_i) = \left(\prod_{u\in in(s_i)} \sum_{v\in in(t_j)} P(v|u)\right)^{\frac{1}{|in(s_i)|}}$$

where $P(v|u)$ refers to the lexical translation probability from the source word u to the target word v within the sub-tree spans, while $P(u|v)$ refers to that from target to source; $in(s_i)$ refers to the word set for the internal span of the source sub-tree s_i, while $in(t_j)$ refers to that of the target sub-tree t_j.

4.2 Internal-External Lexical Features

Intuitively, lexical translation probabilities tend to be high within the translational equivalence, while low within the non-equivalent counterparts. According to this, we define the internal-external lexical feature functions as follows:

$$\varphi(s_i|t_j) = \frac{\sum_{v\in in(t_j)} \max_{u\in out(s_i)} \left\{(P(u|v)\cdot P(v|u))^{\frac{1}{2}}\right\}}{|in(t_j)|}$$

$$\varphi(t_j|s_i) = \frac{\sum_{u\in in(s_i)} \max_{v\in out(t_j)} \left\{(P(u|v)\cdot P(v|u))^{\frac{1}{2}}\right\}}{|in(s_i)|}$$

where $out(s_i)$ refers to the word set for the external span of the source sub-tree s_i, while $out(t_j)$ refers to that of the target sub-tree t_j. We choose a representation different from the internal lexical feature scores, since for cases with small inner span and large outer span, the sum of internal-external scores may be overestimated. As a result, we change the *sum* operation into *max*, which is easy to be normalized.

4.3 Internal Word Alignment Features

Although the word alignment information within bilingual sentence pairs is to some extent not reliable, the links of word alignment account much for the co-occurrence of the aligned terms. We define the internal word alignment features as follows:

$$\rho(s_i, t_j) = \frac{\sum_{v\in in(t_j)} \sum_{u\in in(s_i)} \delta(u,v)\cdot (P(u|v)\cdot P(v|u))^{\frac{1}{2}}}{(|in(s_i)|\cdot|in(t_j)|)^{\frac{1}{2}}}$$

where

$$\delta(u,v) = \begin{cases} 1 & \text{if } (u,v) \text{ is aligned} \\ 0 & \text{otherwise} \end{cases}$$

The binary function $\delta(u,v)$ is introduced to trigger the computation only when a word aligned link exists for the two words (u,v) within the sub-tree span.

4.4 Internal-External Word Alignment Features

Similar to lexical features, we also introduce internal-external word alignment features as follows:

$$\sigma(s_i|t_j) = \frac{\sum_{v\in in(t_j)} \sum_{u\in out(s_i)} \delta(u,v)\cdot (P(u|v)\cdot P(v|u))^{\frac{1}{2}}}{(|out(s_i)|\cdot|in(t_j)|)^{\frac{1}{2}}}$$

$$\sigma(t_j|s_i) = \frac{\sum_{v\in out(t_j)} \sum_{u\in in(s_i)} \delta(u,v)\cdot (P(u|v)\cdot P(v|u))^{\frac{1}{2}}}{(|in(s_i)|\cdot|out(t_j)|)^{\frac{1}{2}}}$$

where

$$\delta(u,v) = \begin{cases} 1 & \text{if } (u,v) \text{ is aligned} \\ 0 & \text{otherwise} \end{cases}$$

4.5 Tree Structural Features

In addition to the lexical correspondence, we also capture the structural divergence by introducing the tree structural features as follows:

Span difference: Translational equivalent sub-tree pairs tend to share similar length of spans. Thus the model will penalize the candidate sub-tree pairs with largely different length of spans.

$$\xi(s_i, t_j) = \left|\frac{|in(s_i)|}{\max_{1\le k\le I}(|in(s_k)|)} - \frac{|in(t_j)|}{\max_{1\le h\le J}(|in(t_h)|)}\right|$$

Number of Descendants: Similarly, the number of the root's descendants of the aligned sub-trees should also correspond.

$$\tau(s_i, t_j) = \left|\frac{|R(s_i)|}{\max_{1\le k\le I}(|R(s_k)|)} - \frac{|R(t_j)|}{\max_{1\le h\le J}(|R(t_h)|)}\right|$$

where $R(.)$ refers to the descendant set of the root to an individual sub-tree.

Tree Depth difference: Intuitively, translational equivalent sub-tree pairs tend to have similar depth from the root node of the parse tree. We can further allow the model to penalize the candidate sub-tree pairs with different distance from the root node.

$$\omega(s_i, t_j) = \left| \frac{Depth(s_i)}{Height(S^I)} - \frac{Depth(t_j)}{Height(T^J)} \right|$$

4.6 Binary Grammatical Features

In the previous sections, we design some score based feature functions to describe syntactic tree structural similarities, rather than directly using the substructures. This is because for limited annotated tree alignment data, features like tokens and grammar rules are rather sparse. In spite of this, we still have a closed set of grammatical tags which can be covered by a small amount of data. Therefore, we use the combination of root grammar tags of the sub-tree pairs as binary features.

5 Training

We train the sub-tree alignment model in two steps:

Firstly, we learn the various feature functions. On one hand, GIZA++ is offline trained on a large amount of bilingual sentences to compute the lexical and word alignment features. On the other hand, the tree structural features, similar to word and phrase penalty features in phrase based SMT models, are computed online for both training and testing.

Secondly, we train the MaxEnt model in Eq. 1, using the training corpus which consists of the bilingual parse tree pairs with manually annotated sub-tree alignment. We apply the widely used GIS (Generalized Iterative Scaling) algorithm (Darroch and Ratcliff, 1972) to optimize λ_1^M. In practice, we modify Och's implementation YASMET.

Since we consider each sub-tree pair as an individual instance, it is easy to see that the negative samples heavily overwhelm the positive ones. For GIS training, such a skewed distribution easily drives the parameters to facilitate the negative instances. We address this problem by giving more weight to the positive training instances.

6 Experiments on Sub-Tree Alignments

We utilize two different corpora to evaluate the proposed sub-tree alignment method and its capability to plug in the related applications respectively. One is HIT English Chinese parallel tree bank with both tree structure and sub-tree alignment manually annotated. The other is the automatically parsed bilingual tree pairs (allowing minor parsing errors) with manually annotated sub-tree alignment. The latter benefits MT task, since most linguistically motivated syntax SMT systems require a held-out automatic parser to achieve rule induction.

6.1 Data preparation

For the gold standard corpus based experiment, we use HIT [1] Chinese-English parallel tree bank, which is collected from English learning text books in China as well as example sentences in dictionaries. It consists of 16131 gold standard parse tree pairs with manually annotated sub-tree alignments. The annotation strictly preserves the semantic equivalence, i.e., it only conducts *sure* links in the internal node level, while ignoring *possible* links adopted in word alignment. In contrast, in the POS level, n-to-n links are allowed in annotation. In order to be consistent with the definition in Section 2, we delete those n-to-n links in POS level. The word segmentation, tokenization and parse-tree in the corpus are manually constructed or checked. The Chinese parse tree in HIT tree bank adopts a different annotation criterion from the Penn TreeBank annotation, which is designed by the HIT research team. The new criterion can better facilitate the description of some rare structural phenomena in Chinese. The English parse tree still uses Penn TreeBank annotation. The statistics of HIT corpus is shown in Table 1.

	Chinese	English
# of Sentence pair	16131	
Avg. Sentence Length	13.06	13.00
Avg. # of sub-tree	21.60	23.74
Avg. # of alignment	11.71	

Table 1. Statistics for HIT gold standard Tree bank

Since the induction of sub-tree alignment is designed to benefit the machine translation modeling, it is preferable to conduct the sub-tree alignment experiment on the corpus for MT evaluation. However, most syntax based SMT systems use an automatic parser to facilitate training and decoding, which introduces parsing errors. Additionally, the gold standard HIT corpus is not applicable for MT

[1] HIT corpus is designed and constructed by HIT mitlab. http://mitlab.hit.edu.cn/index.php/resources.html . We licensed the corpus from them for research usage.

experiment due to problems of domain divergence, annotation discrepancy (Chinese parse tree adopts a different grammar from Penn Treebank annotations) and degree of tolerance for parsing errors.

Due to the above issues, we annotate a new data set to apply the sub-tree alignment in machine translation. We randomly select 300 bilingual sentence pairs from the Chinese-English FBIS corpus with the length ≤ 30 in both the source and target sides. The selected plain sentence pairs are further parsed by Stanford parser (Klein and Manning, 2003) on both the English and Chinese sides. We manually annotate the sub-tree alignment for the automatically parsed tree pairs according to the definition in Section 2. To be fully consistent with the definition, we strictly preserve the semantic equivalence for the aligned sub-trees to keep a high precision. In other words, we do not conduct any doubtful links. The corpus is further divided into 200 aligned tree pairs for training and 100 for testing. Some initial statistic of the automatically parsed corpus is shown in Table 2.

		Chinese	English
Train	# of Sentence pair	200	
	Avg. Sentence Length	17	20.84
	Avg. # of sub-tree	28.87	34.54
	Avg. # of alignment	17.07	
Test	# of Sentence pair	100	
	Avg. Sentence Length	16.84	20.75
	Avg. # of sub-tree	29.18	34.1
	Avg. # of alignment	17.75	

Table 2. FBIS selected Corpus Statistics

6.2 Baseline approach

We implement the work in Tinsley et al. (2007) as our baseline methodology.

Given a tree pair $< S^I, T^J >$, the baseline approach first takes all the links between the sub-tree pairs as alignment hypotheses, i.e., the Cartesian product of the two sub-tree sets:

$$\{s_1, \dots, s_i, \dots, s_I\} \times \{t_1, \dots, t_j, \dots, t_J\}$$

By using the lexical translation probabilities, each hypothesis is assigned an alignment score. All hypotheses with zero score are pruned out. Then the algorithm iteratively selects the link of the sub-tree pairs with the maximum score as a *sure* link, and blocks all hypotheses that contradict with this link and itself, until no non-blocked hypotheses remain.

The baseline system uses many heuristics in searching the optimal solutions with alternative score functions. Heuristic *skip1* skips the tied hypotheses with the same score, until it finds the highest-scoring hypothesis with no competitors of the same score. Heuristic *skip2* deals with the same problem. Initially, it skips over the tied hypotheses. When a hypothesis sub-tree pair (s_i, t_j) without any competitor of the same score is found, where neither s_i nor t_j has been skipped over, the hypothesis is chosen as a *sure* link. Heuristic *span1* postpones the selection of the hypotheses on the POS level. Since the highest-scoring hypotheses tend to appear on the leaf nodes, it may introduce ambiguity when conducting the alignment for a POS node whose child word appears twice in a sentence.

The baseline method proposes two score functions based on the lexical translation probability. They also compute the score function by splitting the tree into the internal and external components.

Tinsley et al. (2007) adopt the lexical translation probabilities dumped by GIZA++ (Och and Ney, 2003) to compute the span based scores for each pair of sub-trees. Although all of their heuristics combinations are re-implemented in our study, we only present the best result among them with the highest Recall and F-value as our baseline, denoted as *skip2_s1_span1*[2].

6.3 Experimental settings

- To examine the effectiveness of the proposed features, we

(1) learn the word alignment using the combination of the 14k of HIT tree bank and FBIS (240k) corpus for both our approach and the baseline method, and divide the remaining HIT corpus as 1k for training and 1k for testing.

(2) learn the word alignment on the entire FBIS training corpus (240k) for both our approach and the baseline method. We then train and test on FBIS corpus of 200 and 100 respectively as stated in Table 2.

- In our task, annotating large amount of sub-tree alignment corpus is time consuming and more difficult compared with the tasks like sequence labeling. One of the important issues we are concerned about is whether we can achieve an acceptable performance with limited training data. We

(3) adopt the entire FBIS data (240k) to learn the word alignment and various amount of HIT gold standard corpus to train the MaxEnt model. Then we test the alignment performance on the same HIT test set (1k) as (1).

[2] s1 denotes score function 1 in Tinsley et al. (2007)

Features	Precision	Recall	F-value
In Lexical	50.96	48.11	49.49
+ InOut Lexical	55.26	53.84	54.54
+ In word align	56.16	60.59	58.29
+ InOut word align	55.80	62.25	58.85
+ Tree Structure	57.64	63.11	60.25
+ Binary Feature	73.14	85.11	78.67
Baseline [Tinsley 2007]	64.14	66.99	65.53

Table 3. Sub-tree alignment of different feature combination for HIT gold standard test set

Features	Precision	Recall	F-value
In Lexical	63.53	54.87	58.88
+ InOut Lexical	66.00	63.66	64.81
+ In word align	70.89	75.88	73.30
+ InOut word align	72.05	80.16	75.89
+ Tree Structure	72.03	80.95	76.23
+ Binary Feature	76.08	85.29	80.42
Baseline [Tinsley 2007]	70.48	78.70	74.36

Table 4. Sub-tree alignment of different feature combination for FBIS test set

• We further test the robustness of our method under different amount of data to learn the lexical and word alignment feature functions. We gradually change the amount of FBIS corpus to train the word alignment. Then we

(4) use the same training (1k) and testing data (1k) with (1);

(5) use FBIS corpus 200 to train MaxEnt model and 100 for testing similar to (2).

6.4 Experimental results

We use Precision, Recall and F-score to measure the alignment performance and obtain the results as follows:

• In Table 3 and 4 for **Exp (1)** and **(2)** respectively, we show that by incrementally adding new features in a certain order, the F-value consistently increases and both outperform the baseline method. From both tables, we find that the **Binary features**, with the combination of root grammar tags of the sub-tree pairs, significantly improve the alignment performance. We also try the different combinations of the parent, child or even siblings to the root nodes. However, all these derivative configurations decrease the performance. We attribute the ineffectiveness to data sparseness. Further exploration suggests that the binary feature in HIT gold standard corpus exhibits a substantially larger improvement against other features than FBIS corpus (Table 3 against Table 4). The reason could be that the grammar tags in the gold standard corpus are accurate, while FBIS corpus suffers from parsing errors. Apart from that, the lexical/word-alignment features in Table 3 do not perform well, since the word alignment is trained mainly on the cross domain FBIS corpus. This is also an important reason why there is a large gap in performance between Table 3 and 4, where the automatic parsed FBIS corpus performs better than HIT gold standard tree bank in all configurations as well as the baseline.

• In Fig. 2(a) for **Exp (3)**, we examine performance under different amount of training data from 1k to 15k. The results change very little with over the amount of 1k. Even with only 0.25k training data, we are able to gain a result close to the best performance. This suggests that by utilizing only a small amount of sub-tree aligned corpus, we can still achieve a satisfactory alignment result. The benefits come from the usage of the score based feature functions by avoiding using sub-structures as binary features, which suffers from the data sparseness problem.

• In Fig. 2(b-e) for **Exp (4&5)**, we find that increasing the amount of corpus to train GIZA++ does not improve much for the proposed method on both HIT gold standard corpus (Fig. 2: b, c) and the automatic parsed data (Fig. 2: d, e). This is due to the various kinds of features utilized by the MaxEnt model, which does not bet on the lexical and word alignment feature too much. As for the baseline method, we can only detect a relatively large improvement in the initial increment of corpus, while later additions do not help. This result suggests that the baseline method is relatively less extensible since it works completely on the lexical similarities which can be only learned from the word alignment corpus.

7 Experiments on Machine Translation

In addition to the alignment evaluation, we conduct MT evaluation as well. We explore the effectiveness of sub-tree alignment for both phrase and linguistically motivated syntax based systems.

7.1 Experimental configuration

In the experiments, we train the translation model on FBIS corpus (7.2M (Chinese) + 9.2M (English) words in 240,000 sentence pairs) and train a 4-gram language model on the Xinhua portion of the English Gigaword corpus (181M words) using the SRILM Toolkits (Stolcke, 2002). We use these

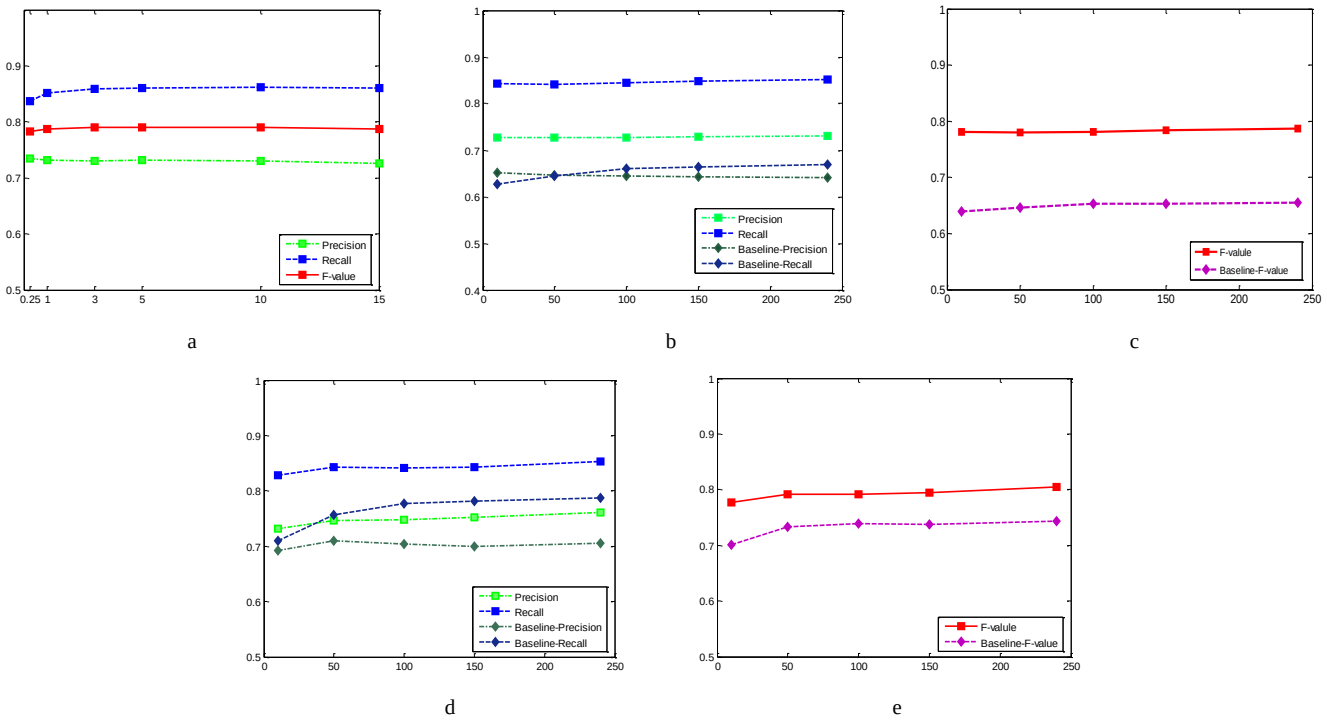

Figure 2: a. Precision/Recall/F-score for various amount of training data (k).
b~e. Various amount of data to train word alignment
b. Precision/Recall for HIT test set. c. F-score for HIT test set.
d. Precision/Recall for FBIS test set. e. F-score for FBIS test set.

sentences with less than 50 characters from the NIST MT-2002 test set as the development set (to speed up tuning for syntax based system) and the NIST MT-2005 test set as our test set. We use the Stanford parser (Klein and Manning, 2003) to parse bilingual sentences on the training set and Chinese sentences on the development and test set. The evaluation metric is case-sensitive BLEU-4.

For the phrase based system, we use Moses (Koehn et al, 2007) with its default settings. For the syntax based system, since sub-tree alignment can directly benefit Tree-2-Tree based systems, we apply the sub-tree alignment in an SMT system based on Synchronous Tree Substitution Grammar (STSG) (Zhang et al., 2007). The STSG based decoder uses a pair of *elementary tree* as a basic translation unit. Recent research on tree based systems shows that relaxing the restriction from tree structure to tree sequence structure (Synchronous Tree Sequence Substitution Grammar: STSSG) significantly improves the translation performance (Zhang et al., 2008). We implement the STSG/STSSG based model in Pisces decoder with the same features and settings in Sun et al. (2009). The STSSG based decoder translates each span iteratively in a bottom up manner which guarantees that when translating a source span, any of its sub-spans has already been translated. The STSG based experiment can be easily achieved by restricting the translation rule set in the STSSG decoder to be elementary tree pairs only.

For the alignment setting of the baselines, we use the word alignment trained on the entire FBIS(240k) corpus by GIZA++ with heuristic grow-diag-final for Moses and the syntax systems and perform rule extraction constrained on the word alignment. As for the experiments adopting sub-tree alignment, we use the above word alignment to learn lexical/word alignment features, and train the sub-tree alignment model with FBIS training data (200).

7.2 Experimental results

Utilizing the syntactic rules only has been argued to be ineffective (Koehn et al., 2003). Therefore, instead of using the sub-tree aligned rules only, we try to improve the word alignment constrained rule set by sub-tree alignment as shown in Table 5.

Firstly, we try to *Directly Concatenate* (DirC) the sub-tree alignment constraint rule set[3] to the original syntax/phrase rule set based on word alignment. Then we re-train the MT model based

[3] For syntax based system, it's just the sub-tree pairs deducted from the sub-tree alignment; for phrase based system, it's the phrases with context equivalent to the aligned sub-tree pairs.

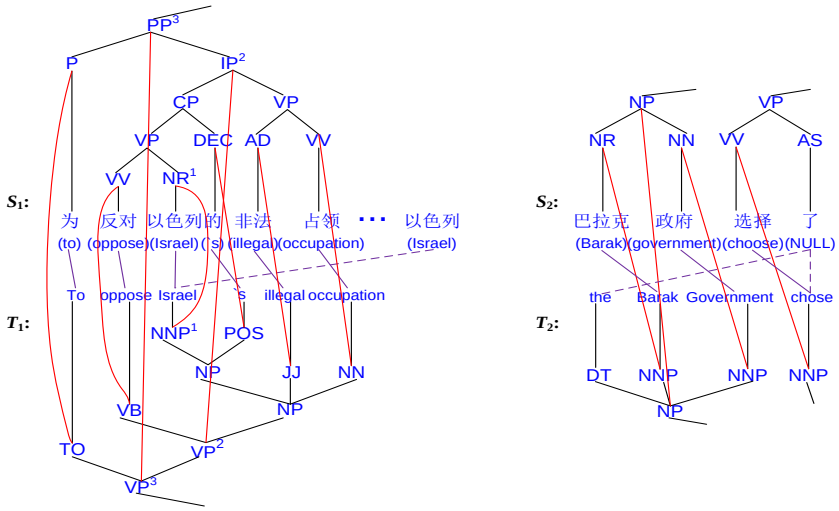

Figure 3: Comparison between Sub-tree alignment results and Word alignment results

on the obtained rule set. Tinsley et al. (2009) attempts different duplication of sub-tree alignment constraint rule set to append to the original phrase rule set and reports positive results. However, as shown in Table 5, we only achieve very minor improvement (in STSSG based model the score even drops) by direct introducing the new rules.

Secondly, we propose a new approach to utilize sub-tree alignment by modifying the rule extraction process. We allow the bilingual phrases which are consistent with *Either* Word alignment *or* Sub-tree alignment (EWoS) instead of to be consistent with word alignment only. The results in Table 5 show that EWoS achieves consistently better performance than the baseline and DirC method. We also find that sub-tree alignment benefits the STSSG based model less compared with other systems. This is probably due to the fact that the STSSG based system relies much on the tree sequence rules.

To benefit intuitive understanding, we provide two alignment snippets in the MT training corpus in Fig. 3, where the red lines across the non-terminal nodes are the sub-tree aligned links conducted by our model, while the purple lines across the terminal nodes are the word alignment links trained by GIZA++. In the first example, the word *Israel* is wrongly aligned to two "以色列"s by GIZA++, where the wrong link is denoted by the dash line. This is common, since in a compound sentence in English, the entities appeared more than once are often replaced by pronouns at its later appearances. Therefore, the syntactic rules constraint by NR^1-NNP^1, IP^2-VP^2 and PP^3-VP^3 respectively cannot be extracted for syntax systems; while for phrase systems, context around the first "以色列" cannot be fully explored. In the second example, the empty word "了" is wrongly aligned, which usually occurs in Chinese-English word alignment. As shown in Fig. 3, both cases can be resolved by sub-tree alignment conducted by our model, indicating that sub-tree alignment is a decent supplement to the word alignment rule set.

System	Rules	BLEU
Moses	BP*	23.86
	DirC	24.12
	EWoS	24.45
Syntax STSG	STSG	24.71
	DirC	24.91
	EWoS	25.21
Syntax STSSG	STSSG	25.92
	DirC	25.88
	EWoS	26.12

Table 5. MT evaluation on various systems
BP* denotes bilingual phrases.
BP, STSG, STSSG are baseline rule sets using word alignment to constrain rule extraction.

8 Conclusion

In this paper, we propose a framework for bilingual sub-tree alignment using Maximum Entropy model. We explore various lexical and structural features to improve the alignment performance. We also manually annotated the automatic parsed tree pairs for both alignment evaluation and MT experiment. Experimental results show that our alignment framework significantly outperforms the baseline method and the proposed features are very effective to capture the bilingual structural similarity. Additionally, we find that our approach can perform well using only a small amount of sub-tree aligned training corpus. Further experiment shows that our approach benefits both phrase and syntax based MT systems.

References

David Burkett and Dan Klein. 2008. *Two languages are better than one* (*for syntactic parsing*). In Proceedings of EMNLP-08. 877-886.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang and Ignacio Thayer. 2006. *Scalable Inference and training of context-rich syntactic translation models*. In Proceedings of COLING-ACL-06. 961-968.

Jonathan Graehl and Kevin Knight. 2004. *Training tree transducers*. In Proceedings of HLT-NAACL-2004. 105-112.

Declan Groves, Mary Hearne, and Andy Way. 2004. *Robust sub-sentential alignment of phrase-structure trees*. In Proceedings of COLING-04, pages 1072-1078.

Liang Huang, Kevin Knight and Aravind Joshi. 2006. *Statistical syntax-directed translation with extended domain of Locality*. In Proceedings of AMTA-06.

Kenji Imamura. 2001. *Hierarchical Phrase Alignment Harmonized with Parsing*. In Proceedings of NLPRS. 377-384.

Dan Klein and Christopher D. Manning. 2003. *Accurate Unlexicalized Parsing*. In Proceedings of ACL-03. 423-430.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin and Evan Herbst. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*. In Proceedings of ACL-07. 177-180.

Philipp Koehn, Franz Josef Och and Daniel Marcu. 2003. *Statistical Phrase-based Translation*. In Proceedings of HLT-NAACL-2003. 48-54.

Yang Liu, Qun Liu and Shouxun Lin. 2006. *Tree-to-String alignment template for statistical machine translation*. In Proceedings of ACL-06, 609-616.

Daniel Marcu, Wei Wang, Abdessamad Echihabi and Kevin Knight. 2006. *SPMT: statistical machine translation with syntactified target language phrases*. In Proceedings of EMNLP-06. 44-52.

Franz Josef Och and Hermann Ney. 2003. *A systematic comparison of various statistical alignment models*. Computational Linguistics, 29(1):19-51, March.

Andreas Stolcke. 2002. *SRILM - an extensible language modeling toolkit*. In Proceedings of ICSLP-02. 901-904.

Jun Sun, Min Zhang and Chew Lim Tan. 2009. *A non-contiguous Tree Sequence Alignment-based Model for Statistical Machine Translation*. In Proceedings of ACL-IJCNLP-09. 914-922.

John Tinsley, Ventsislav Zhechev, Mary Hearne, and Andy Way. 2007. *Robust language pair-independent sub-tree alignment*. In Proceedings of Machine Translation Summit-XI-07.

John Tinsley, Mary Hearne, and Andy Way. 2009. *Parallel treebanks in phrase-based statistical machine translation*. In Proceedings of CICLING-09.

Min Zhang, Hongfei Jiang, AiTi Aw, Jun Sun, Sheng Li and Chew Lim Tan. 2007. *A tree-to-tree alignment-based model for statistical machine translation*. In Proceedings of MT Summit-XI -07. 535-542.

Min Zhang, Hongfei Jiang, AiTi Aw, Haizhou Li, Chew Lim Tan and Sheng Li. 2008. *A tree sequence alignment-based tree-to-tree translation model*. In Proceedings of ACL-08. 559-567.