Improving Word Alignment Based on Extended Inversion Transduction Grammar

Chung-Chi Huang†
CLCLP, TIGP, Academia Sinica, Taiwan
u901571@gmail.com

Wei-Teh Chen
ISA, NTHU, Taiwan
weitehchen@gmail.com

Jason S. Chang
CS, NTHU, Taiwan
jason.jschang@gmail.com

Abstract

We propose a fusion of Inversion Transduction Grammar model with IBM-style notation of fertility to improve word-aligning performance. In our approach, binary context-free grammar rules on the source language, accompanied with orientation preferences on the target, and fertilities of words are leveraged to construct a syntax-based statistical translation model. Our model, inherently possessing the characteristic of ITG restrictions and allowing for many consecutive words aligned to one and vise versa, outperforms original ITG model and GIZA++ not only in alignment error rate (23% and 14% error reduction) but in consistent phrase error rate (13% and 9% error reduction) as well. Better performance in these two evaluation metrics will lead to better phrase-based machine translation with great possibility.

Keywords

Word alignment, inversion transduction grammar, IBM models, alignment error rate, parsing, and GIZA++.

1. Introduction

Statistical translation model is a model which detects word correspondences within sentence pairs whether relied on lexical information or syntactic aspects of languages involved. In spite of the fact that the methodologies varies, the intention is clear—trying to obtain better word alignment results since a better translation model implies better performance in various linguistic applications. Among them are phrase-based machine translation (Och and Ney, 2004; David Chiang, 2005; Liu et al., 2006) and inference of syntactic translation rules (Galley et al., 2004; Galley et al., 2006).

Since the pioneering work of (Brown et al., 1988), there have been a myriad of subsequent researches related to statistical translation model. They could mainly be classified into two categories: one paying little attention to the grammars of the languages (Vogel et al., 1996; Och and Ney, 2000; Toutanova et al., 2002) and the other explicitly utilizing languages' structural or syntactic information (Wu, 1997; Yamada and Knight, 2001; Cherry and Lin, 2003; Gildea, 2004; Zhang and Gildea, 2005). With more and more accurate syntactic analyzers (such as part-of-speech tagger and Stanford parser) being developed and in view of the deficiency in modeling grammatical facets of languages IBM-like models experience, latter researches have received increasing attention.

To incorporate syntax of involved languages, Yamada and Knight (2001) accepted source-language (SL, such as English) parse trees as input and made use of reordering, inserting and translating operations to transform the input parse trees into counterpart target-language (TL, such as French) strings. In contrast to flattening the input parse trees to do the transformation (reordering, inserting and translating) for every node, Wu's ITG (1997) attempted to associate each production rule commonly shared by two languages with word orientation. Besides, instead of accepting parse trees produced by a monolingual parser, Wu's approach makes possible constructing bilingual parse trees synchronously.

The strengths of two models are discussed in (Zhang and Gildea, 2004), which also found data-oriented bilingual parsing turned out to outperform tree-to-string model for word-level alignment. Nonetheless, in (Wu, 1997), constituent categories are not differentiated and the probabilities of the *straight* or *inverted* orientation of binary production rules, rather than trained on real-life cases, are all assigned constant.

Inspired by (Zhang et al. 2006), which suggests binarization of synchronous rules improves both speed and accuracy of a syntax-based machine translation system, in this paper, to capture the systematic differences in languages' grammars, such as SVO (English or Chinese), SOV (Japanese) and VSO (Arabic) word orders, we attach the information of identical or dissimilar orientation of languages' counterparts onto binary SL CFG rules, resulting in grammatical rewrite rules biased on SL side, or more specifically, biased ITG rules, *bITG* for short. For instance, the similar VO construct in both English and Chinese can be observed from the high probability of the bITG rule $VP \rightarrow [VP\ NP]$ where square bracket indicates the same ordering (*straight*) of the two right-hand-side constituents in both languages when expanding the left-hand-side symbol. On the contrary, the different VO construct in English and Japanese can be modeled using high *inverted* probability of bITG rule $VP \rightarrow \langle VP\ NP \rangle$ where pointed bracket denotes we expand the left-hand-side label into two right-hand-side symbols in reverse orientation in two languages. However, both bITG rules are inferred from the same binary CFG rule ($VP \rightarrow VP\ NP$) of the source language, English, only with different order preferences on the target end.

Furthermore, in our model fusing bITG model with IBM-style fertilities, many contiguous words on the source can be aligned to one word on the target and vice versa based on fertility probabilities of words. Originally, Wu’s ITG (1997) only allowed for, at most, one-to-one word alignment, which may decrease the accuracy of the bilingual parse trees and, in turn, the performance on word alignments. This one-to-one restriction on word-aligning is especially not suitable for language pair like English and Chinese since the tokenization work of Chinese sentences prior to word alignment would introduce many many-to-one or one-to-many links in that the resulting segments in Chinese sentences are independent of words on English side. That is, the segmentations in Chinese can be under- or over- segmented for the corresponding words in English. As a result, the translation model accommodating more than one-to-one correspondences is of great importance, especially for such language pair.

Section 2 and 3 describe our model in detail. Section 4 shows experimental results. Discussions are made before conclusion in section 6.

2. The Model

2.1 An Example

First, an example of how bITG rules are exploited to assist in word-aligning sentence pairs is introduced. A more formal description of our model will be discussed in sequent sections.

We assume a parallel sentence pair and POS information of the SL sentence are fed into our model and it, using not only lexical translation rules but the binary SL CFG rules accompanied with orientation preferences of counterparts on the TL, synchronously parses the bilingual sentence pair and yields the word alignments at the leaf level of the bilingual parse tree.

The model assigns probabilities to substring pairs of the bilingual sentences after each of them is associated with possible syntactic labels on the source side. Take the sentence pair and its parse in Figure 1, where spaces in the Chinese sentence are used to distinguish the boundaries of segments and ∗ denotes the *inverted* orientation of the node’s children on the target, for example. The substring pair (positive role, 積極 作用) associated with constituent category *NP* will be assigned a probability. In this particular parse, the best probability of parsing (positive role, 積極 作用) is the product of probabilities of *straight* bITG rule, $NP \rightarrow [JJ\ NN]$, and lexical translation rules, $JJ \rightarrow$ positive/積極 and $NN \rightarrow$ role/作用 where / denotes word correspondence in both languages. The higher probability of the rule $NP \rightarrow [JJ\ NN]$ than that of the *inverted* rule $NP \rightarrow \langle JJ\ NN \rangle$ not just instructs the model to align the right-hand-side counterparts of two languages in a *straight* fashion more, but implies the similar word orientation for the syntactic structure in English and Chinese.

On the other hand, we would notice that the beginning half “*These factors will continue to play a positive role*” is translated into the back of the Chinese sentence whereas the ending half “*after its return*” is translated into the beginning. This phenomenon is very common while translating one language into another. The *inverted* word order rules trained on parallel corpus, like $S \rightarrow \langle S\ PP \rangle$, are devised to capture the systematic differences of the languages’ grammars.

In the end, taking into account both the probabilities of lexical and grammatical rewrite rules and fertilities of words in languages, the model endeavors to find the best parse that applies more appropriate production rules to match the similarities and dissimilarities of two languages, which, in turn, yields better word alignment results. As for this example parse, the sentence pair associated with the syntactic label *S* results in best bilingual parse tree whose probability is estimated by the product of probabilities of the bITG rules, $S \rightarrow \langle S\ PP \rangle$, and root’s two children, (*These factors will continue to play a positive role*, 這些 條件 將會 繼續 發揮 積極 作用)$_S$ and (*after its return*, 香港 回歸 後)$_{PP}$.

We actually obtain probabilities of bITG rules, consisting of lexical rules and binary SL CFG rules with word orientation preferences on the target, and fertilities of words from a parallel corpus and SL CFG. Section 3 describes the training algorithm.

2.2 Runtime Parsing

In this section, we extend Wu’s ITG (1997) such that our model incorporates the grammatical constituents on the source language and accommodates the cases of many contiguous words on the source aligned to one on the target and vice versa.

The English-French notation is used throughout this paper. *E* and *F* denote the source and target language respectively and e_i stands for the i-th word in sentence *e* in language *E* and f_j for the j-th word in sentence *f* in *F*.

As mentioned in (Wu, 1997; Zens and Ney, 2003), the ITG constraint allows for a polynomial-time parsing algorithm, based on a recursion equation that can be resolved by a CYK-style parser. During a parse of a sentence pair in our model, a table of $\delta_{p,s,t,u,v}$, which represents the *best* probability of parsing substring pair $(e_{s+1} \cdots e_t, f_{u+1} \cdots f_v)$

English sentence: These factors will continue to play a positive role after its return

Chinese sentence: 香港 回歸 後 這些 條件 將會 繼續 發揮 積極 作用

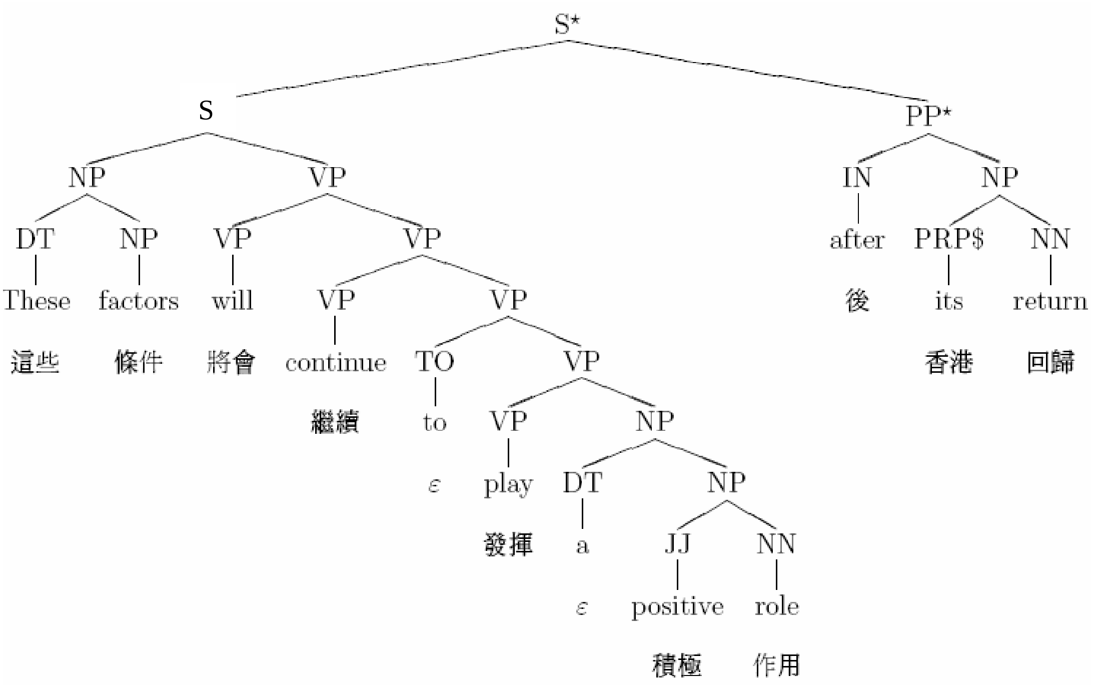

Figure 1. An example sentence pair and its bilingual parse tree

related to a syntactic label *p* on the E side, is constructed. We initialize this table with probabilities of one-to-one, one-to-zero and zero-to-one word correspondences limited on the scope of the sentence pair. Afterwards, relied on the work done previously, many-to-many word correspondences and parsing results of longer substring pairs would unveil themselves in a bottom-top manner. Meanwhile, integration of fertilities of words into the model further boosts the word-aligning performance.

Following is the CYK parsing algorithm in our model, where we parse a sentence pair (e, f), $(e_1 \cdots e_m, f_1 \cdots f_n)$, and the POS tag sequence of *e* is $(t_1, \cdots, t_m)$. In the algorithm, $\mathrm{P}(L \to t)$ denotes probability of a lexical rule and *t* could be e_i / f_j, e_i / ε and ε / f_j where ε stands for NULL, while $\mathrm{P}(L \to [R_1\ R_2])$ and $\mathrm{P}(L \to \langle R_1\ R_2 \rangle)$ denote probabilities of binary bITG rules where R_1 and R_2 indicate the right-hand-side syntactic constituents of the CFG rules in *E*. Furthermore, $\Pr(\Phi_{e_i} = x)$ and $\Pr(\Phi_{f_j} = x)$ represent the probabilities of fertilities of e_i and f_j being associated with *x*, respectively.

Parsing Algorithm

1. Initial Step

For $1 \le i \le m, 1 \le j \le n$

$$\delta_{t_i, i-1, i, j-1, j} = \mathrm{P}\left(t_i \to e_i / f_j\right) \times \Pr\left(\Phi_{e_i} = 1\right) \times \Pr\left(\Phi_{f_j} = 1\right)$$

For every $L \to t_i \in$ grammar rules in E

$$\delta_{L, i-1, i, j-1, j} = \mathrm{P}\left(L \to e_i / f_j\right) \times \Pr\left(\Phi_{e_i} = 1\right) \times \Pr\left(\Phi_{f_j} = 1\right)$$

For $1 \le i \le m, 0 \le j \le n$

$$\delta_{t_i, i-1, i, j, j} = \mathrm{P}\left(t_i \to e_i / \varepsilon\right) \times \Pr\left(\Phi_{e_i} = 0\right)$$

For every $L \to t_i \in$ grammar rules in E

$$\delta_{L, i-1, i, j, j} = \mathrm{P}\left(L \to e_i / \varepsilon\right) \times \Pr\left(\Phi_{e_i} = 0\right)$$

For $0 \le i \le m, 1 \le j \le n, L \in$ syntactic labels in E

$$\delta_{L, i, i, j-1, j} = \mathrm{P}\left(L \to \varepsilon / f_j\right) \times \Pr\left(\Phi_{f_j} = 0\right)$$

2. Recurrent Step

$$\delta_{p,s,t,u,v} = \max_{\substack{q,r\in\text{syntax labels on } E\\ s\le s'\le t\\ u\le u'\le v}} \left\{ \begin{array}{l} \mathrm{P}\left(p\rightarrow[q\ r]\right)\times\delta_{q,s,s',u,u'}\times\delta_{r,s',t,u',v}, \\ \mathrm{P}\left(p\rightarrow\langle q\ r\rangle\right)\times\delta_{q,s,s',u',v}\times\delta_{r,s',t,u,u'} \end{array} \right\}$$

However, for $\delta_{p,s,t,u-1,u}$, the possible choice to parsing the substring pair also includes $\Pr\left(\Phi_{f_u}=(t-s)\right)\times$

$$\max_{\substack{q,r\in\text{syntax}\\ \text{labels on } E}} \left\{ \mathrm{P}\left(p\rightarrow[q\ r]\right)\times\frac{\delta_{q,s,s+1,u-1,u}}{\Pr\left(\Phi_{f_u}=1\right)}\times\frac{\delta_{r,s+1,t,u-1,u}}{\Pr\left(\Phi_{f_u}=(t-s-1)\right)} \right\}.$$

Similar principle applies for $\delta_{p,s-1,s,u,v}$.

2.3 Pruning

Although the complexity of described algorithm is polynomial-time, the execution time grows rapidly with the increase in the variety of syntactic labels, from three structural labels in (Wu, 1997) to the syntactic categories of the source language's grammar. As a result, pruning techniques are essential to reduce the time spent on parsing.

We adopt pruning in following two manners. The idea of the first pruning technique is to only keep parse trees whose probabilities fall within the best $N\times\alpha$, where N is the number of possible parses for SL substring $e_{s+1}\cdots e_t$ and a constant length of the TL substring, and α is a real number between 0 and 1. In other words, we remove less probable parse trees that are not in the best $N\times\alpha$ ones.

The second pruning technique is related to the ratio of the length of SL and TL substring. $\delta_{p,s,t,u,v}$ will be removed, or not calculated, if $(t-s)/(v-u)$ is smaller than θ_{ratio} or larger than $1/\theta_{ratio}$ where $0\le\theta_{ratio}\le 1$, since few words will be aligned to more than $1/\theta_{ratio}$ words in another language. Applying these pruning techniques affects little in the word alignment quality with computational overhead reduced significantly.

3. Probability Estimation

In the first stage of our probabilistic inference process, a word-aligning strategy is applied to acquire the initial word alignments from a sentence-aligned corpus. Thereafter, for every substring pair of each bilingual sentence pair, the SL substring will be related to some possible binary SL CFG rules and, based on initial word alignments, right-hand-side constituents of these rules will be associated with an orientation on the target end. Ultimately, we exploit occurrence of detected bITG rules to estimate probabilities.

3.1 Representation

By applying any existing word-level alignment method, the initial word alignment set **A** for parallel corpus **C** is obtained. **A** is comprised of elements of the form $\left(r, e_{i_1}^{i_2}, f_{j_1}^{j_2}, L, rhs, rel\right)$, which represents substring pair $\left(e_{i_1}\cdots e_{i_2}, f_{j_1}\cdots f_{j_2}\right)$ in sentence pair r has $L\rightarrow rhs$ as the derivation leading to the bilingual structure in the parse tree and rel, either *straight* or *inverted*, as the cross-language word order relations of constituents of rhs, denoting either a sequence of syntactic labels or a single terminating bilingual word pair.

Take the parse in Figure 1 for example, (after its return, 香港 回歸 後)$_{pp}$ would be represented by the 6-tuple $\left(193, e_{10}^{12}, f_1^{3}, PP, IN\ NP, Inverted\right)$ where 193 is the sentence number of this pair, in the word alignment set **A**.

3.2 Training Algorithm

The algorithm starts with a set **H** initialized with the initial word alignment set **A**. Then recursively select two elements, which have not yet been paired up, from **H**. If these two elements have contiguous word sequence on *e* side and exhibit *straight* or *inverted* relation between *e* and *f* based on word alignments, a new tuple representing these two will be added into **H**. At last, we utilize the occurrence in **H** to infer probabilities of bITG rules, $\mathrm{P}\left(L\rightarrow[R_1\ R_2]\right)$, $\mathrm{P}\left(L\rightarrow\langle R_1\ R_2\rangle\right)$ and $\mathrm{P}\left(L\rightarrow t\right)$. Besides, fertility probabilities related to words in both languages are calculated in this algorithm as well.

In the following algorithm, **G** stands for the set of the binary SL CFG rules, $|\mathbf{W}|$ for the number of entries in set **W**, $\mathrm{count}(p;\mathbf{Q})$ for the occurrence of p in set **Q**, δ, a positive integer, for the tolerance of cross-language *straight*/*inverted* word order phenomenon, and Φ_{e_i} and Φ_{f_j} for fertility of the word e_i and f_j, respectively.

Algorithm for Probabilistic Estimation

```
$\mathbf{H}=\mathbf{A}$
For $\left(r,e_{i_1}^{i_2},f_{j_1}^{j_2},L,rhs,rel\right)\in\mathbf{H},\left(r,e_{\bar{i_1}}^{\bar{i_2}},f_{\bar{j_1}}^{\bar{j_2}},\bar{L},\overline{rhs},\overline{rel}\right)\in\mathbf{H}$
have not yet been considered
  If $\left(i_2=\bar{i_1}-1\right)$
    for every $L'\rightarrow L\ \bar{L}\in\mathbf{G}$
      If $\left(j_2+1\le\bar{j_1}\le j_2+\delta\right)$
        $\mathbf{H}=\mathbf{H}\cup\left\{\left(r,e_{i_1}^{\bar{i_2}},f_{j_1}^{\bar{j_2}},L',L\ \bar{L},\text{Straight}\right)\right\}$
      If $\left(\bar{j_2}+1\le j_1\le\bar{j_2}+\delta\right)$
        $\mathbf{H}=\mathbf{H}\cup\left\{\left(r,e_{i_1}^{\bar{i_2}},f_{\bar{j_1}}^{j_2},L',L\ \bar{L},\text{Inverted}\right)\right\}$
  same principle applies when $\bar{i_2}=i_1-1$
```

```
Incorporate words aligned to null, each of which is
denoted using 6-tuple representation, in both
languages into $\mathbf{H}$
For $\left(r, e_{i_1}^{i_2}, f_{j_1}^{j_2}, L, rhs, rel\right) \in \mathbf{H}$
  If $(rhs \neq t)$
    $\mathrm{P}(L \to [R_1\ R_2]) = \frac{\text{count}((*,*,*,L,R_1\ R_2,\text{Straight});\mathbf{H})}{|\mathbf{H}|}$
    $\mathrm{P}(L \to \langle R_1\ R_2 \rangle) = \frac{\text{count}((*,*,*,L,R_1\ R_2,\text{Inverted});\mathbf{H})}{|\mathbf{H}|}$
  Else
    $\mathrm{P}(L \to t) = \frac{\text{count}((*,*,*,L,t,*);\mathbf{H})}{|\mathbf{H}|}$
Based on $\mathbf{A}$ and $\mathbf{C}$, Calculate $\Pr(\Phi_{e_i})$ and $\Pr(\Phi_{f_j})$
using relative frequency
```

4. Experiments

To experiment, we trained our model on a large English-Chinese parallel corpus. To evaluate performance, we examined alignments produced by the proposed model using the evaluation metrics proposed by Och and Ney (2000). For comparison, we also trained GIZA++, a state-of-the-art word-aligning system, on the same corpus.

4.1 Training

We used the news portion of Hong Kong Parallel Text (Hong Kong news) distributed by Linguistic Data Consortium (LDC) as our sentence-aligned corpus $\mathbf{C}$. The corpus consists of 739,919 English-Chinese sentence pairs. English sentences are considered to be the source while Chinese sentences are the target. SL sentences are tagged and TL sentences are segmented before fed into any word alignment strategy or existing system. The average sentence length is 24.4 words for English and 21.5 words for Chinese. On the other hand, PTB section 23[1] production rules distributed by Andrew B. Clegg made up of our binary SL CFG **G**.

4.2 Evaluation

To evaluate our statistical translation model, 114 sentence pairs were chosen randomly from Hong Kong news as our testing data set. For the sake of execution time, we only selected sentence pairs whose length of English and Chinese sentences does not exceed 15, which cover approximately 40% of sentence pairs in the whole Hong Kong news corpus and where better word-aligning results can be obtained using GIZA++. We used the metrics of alignment error rate (AER) proposed by Och and Ney (2000), in which the quality of a word alignment result $\mathbf{A}$ done by an automatic system is evaluated using

$$precision = \frac{|\mathbf{A} \cap \mathbf{P}|}{|\mathbf{A}|},\ recall = \frac{|\mathbf{A} \cap \mathbf{S}|}{|\mathbf{S}|}\ \text{and}$$

$$AER(\mathbf{S},\mathbf{P};\mathbf{A}) = 1 - \frac{|\mathbf{A} \cap \mathbf{S}| + |\mathbf{A} \cap \mathbf{P}|}{|\mathbf{A}| + |\mathbf{S}|},$$

where $\mathbf{S}$ (sure) is the set whose alignments are not ambiguous and $\mathbf{P}$ (possible) is the set consisting of alignments that might or might not exist $(\mathbf{S} \subseteq \mathbf{P})$. Thus, the human-annotated alignments may contain many-to-one and one-to-many relations.

In the experiment, we used an existing system, GIZA++, as our word-aligning strategy in training procedure. In other words, the initial word alignment set was produced by GIZA++ with default settings. Following table illustrates the experimental results of GIZA++, original ITG model in (Wu, 1997), and our extended ITG biased on English side.

Table 1. Results of test data of different systems

	P	R	AER	F
E to F	**.891**	.385	.459	.537
F to E	.882	.533	.333	.664
Refined	.879	.635	.261	.737
ITG	.844	.610	.290	.708
Our model w/o fertility	.866	.638	.263	.735
Our model w/ fertility	.878	**.692**	**.224**	**.774**

In this table[2], P, R and F stand for precision, recall and F-measure[3] respectively. The performance of E to F (E stands for English and F for Chinese), F to E and refinement of both directions, proposed by Och and Ney (2000), of GIZA++, are shown, and so is that of original ITG, which also trained on the lexical output of GIZA++. The results of our translation model without or with the capability of making many-to-one/one-to-many links are listed in the last two rows.

Compared with ITG model that does not distinguish the constituent categories, our model without fertility probability, allowing for at most one-to-one alignment as the original ITG does, achieved 9% reduction in the alignment error rate. It follows the binary SL CFG rules accompanied with ordering preference of the counterparts on the TL trained on parallel corpus do capture the systematic differences of languages' grammars and impose

[1] http://textmining.cryst.bbk.ac.uk/acl05/

[2] $|\mathbf{S}|/|\mathbf{P}|$ is 85.56%.

[3] Calculated using the formula $2 \times \mathrm{P} \times \mathrm{R} / (\mathrm{P}+\mathrm{R})$.

a more realistic and precise reordering constraints on word aligning for the languages pair.

On the other hand, compared to the refined alignments of both directions GIZA++ produced, our model with fertility, which is quite similar to the refined method that accommodates many-to-many alignment relations, increased the recall by 9% while maintaining high precision and overall, achieved 14% alignment error reduction (increased F-measure by 5%).

5. Discussion

In this section, we examine the learnt similarity (*straight*) and difference (*inverted*) in two languages' grammars in aiding the process of word alignment of our model by means of adjacency feature and cohesion constraint, mentioned in (Cherry and Lin, 2003). Subsequently, to evaluate the possibility of leading to better translation performance of a phrase-based MT model if provided the output of our model, we adopt the recently-proposed metric, consistent phrase error rate (CPER) by (Ayan and Dorr, 2006).

5.1 *Straight/Inverted* Orientation

To evaluate the assistance of *straight* orientation of the rules in alignment process, the accuracy of adjacent alignments made by our model is shown in Table 2 and that of refined results of GIZA++ is illustrated for comparison. An ordering, depending on the position of the English word in the sentence, is imposed in order to examine the feature since alignments must have orders before links exhibiting adjacency feature exist.

Table 2. Examination of adjacent links

	Compared to sure links	Compared to possible links
Refined	.835	.869
Our model w/ fertility	**.863**	**.881**

Further, we examine whether the *inverted* orientation of our binary bITG rules does capture the diversities of two grammars and help to make correct crossing links if necessary, or not. For that purpose, after the acquisition of the dependencies of the source sentences by using Stanford parser, the percentage of links violating cohesion constraint, the rate of mapped dependency tree in Chinese having crossing dependencies, is computed.

Table 3. Percentage of links violating cohesion constraint

	Percentage
Refined	.044
Our model w/ fertility	**.037**

We observed 1% to 3% increase in making correct adjacent alignments in Table 2 while in Table 3, our model achieved 16% reduction in percentage of links violating cohesion constraint. Above statistics indicate that the probabilities related to *straight* and *inverted* word orders of ITG rules biased on SL in our model not only impose a more suitable alignment constraints but capture the grammatical relations in two languages, which overall results in better word alignment quality.

5.2 CPER

According to Ayan and Dorr (2006), the intrinsic evaluation metric of AER examines only the quality of word-level alignments but correlates poorly with MT community-standard metric—BLEU score. As a result, we exploit CPER, correlating better with BLEU, to evaluate alignments in the context of phrase-based MT. Precision, recall and CPER are computed as

$$P = \frac{|P_A \cap P_G|}{|P_A|},\ R = \frac{|P_A \cap P_G|}{|P_G|}, \text{ and}$$

$CPER = 1 - \frac{2 \times P \times R}{P + R}$ if the sets of phrases, P_A and P_G, generated by an alignment *A* and manual alignment *G* respectively, are known.

From Table 4, we notice proposed bITG model with fertility yields lowest CPER, with great chance contributing to higher BLEU if a phrase-based MT system accepts the output of our model.

Table 4. Reports on CPER

	P	R	CPER
E to F	.479	.383	.574
F to E	.544	.518	.470
Refined	.573	.606	.411
ITG	.569	.569	.431
Our model w/o fertility	.598	.597	.402
Our model w/ fertility	**.624**	**.626**	**.375**

6. Conclusion

To combine the strengths of competing models, a thought-provoking fusion of IBM-style fertility notation with syntax-based ITG model is described. In our model, *straight/inverted* binary bITG rules, which bypasses the problem that commonly-shared grammatical rules of two languages are difficult to design manually, are statistically trained and devised to boost the word alignment quality. The proposed bITG model with fertilities reduced AER by 14% to 23% and CPER by 9% to 13% comparing to GIZA++ and Wu's ITG (1997), and lower CPER suggests better translation performance if a phrase-based MT is chained after our word-level alignment output. In this paper, the performance of ITG models trained on large-scale parallel corpus is shown for the first time and the result is inspiring.

7. References

[1] D. S. Jurafsky and J. H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, New Jersey, 2000.

[2] A. B. Clegg and A. Shepherd. 2005. Evaluating and integrating Treebank parsers on a biomedical corpus. In *Association for Computational Linguistics Workshop on software 2005*.

[3] C. Cherry and D. Lin. 2003. A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 88-95.

[4] D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263-270.

[5] M. Galley, M. Hopkins, K. Knight and D. Marcu. 2004. What's in a translation rule? In *Proceedings of HLT/NAACL-04.*

[6] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang and I. Thayer. 2006. In *Proceedings of the 44th Annual Conference of the Association for Computational Linguistics*, pages 961-968.

[7] Y. Liu, Q. Liu and S. Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 44th Annual Conference of the Association for Computational Linguistics,* pages 609-616.

[8] F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00)*, pages 440-447.

[9] F. J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417-449.

[10] K. Toutanova, H. T. Ilhan and C. D. Manning. 2002. Extentions to HMM-based statistical word alignment models. In *Proceedings of the Conference on Empirical Methods in Natural Processing Language*.

[11] S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, volume 2, pages 836-841.

[12] D. Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377-403.

[13] K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Conference of the Association for Computational Linguistics (ACL-01).*

[14] R. Zens and H. Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 144-151.

[15] H. Zhang and D. Gildea. 2004. Syntax-based alignment: supervised or unsupervised? In *Proceedings of the 20th International Conference on Computational Linguistics*.

[16] H. Zhang and D. Gildea. 2005. Stochastic lexicalized inversion transduction grammar for alignment. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 475-482.

[17] H. Zhang, L. Huang, D. Gildea and K. Knight. 2006. Synchronous binarization for machine translation. In *Proceedings of the NAACL-HLT.*