Computational Linguistics and Chinese Language Processing
Vol. 14, No. 1, March 2009, pp. 1-18

Fertility-based Source-Language-biased Inversion Transduction Grammar for Word Alignment

Chung-Chi Huang* **and Jason S. Chang+**

Abstract

We propose a version of Inversion Transduction Grammar (ITG) model with IBM-style notation of fertility to improve word-alignment performance. In our approach, binary context-free grammar rules of the source language, accompanied by orientation preferences of the target language and fertilities of words, are leveraged to construct a syntax-based statistical translation model. Our model, inherently possessing the characteristics of ITG restrictions and allowing for many consecutive words aligned to one and vice-versa, outperforms the Bracketing Transduction Grammar (BTG) model and GIZA++, a state-of-the-art word aligner, not only in alignment error rate (23% and 14% error reduction) but also in consistent phrase error rate (13% and 9% error reduction). Better performance in these two evaluation metrics suggests that, based on our word alignment result, more accurate phrase pairs may be acquired, leading to better machine translation quality.

Keywords: Inversion Transduction Grammar, Syntax-based Statistical Translation Model, Word Alignment.

1. Introduction

A statistical translation model is a model which detects word correspondences within sentence pairs, whether relying on lexical information or on syntactic aspects of the involved languages or both. In spite of the fact that methodologies vary, the intention is clear: to obtain better word alignment results so that a better translation model implies better performance in different linguistic applications. Among the methodologies are phrase-based (Och & Ney, 2004; Chiang, 2005; Liu *et al*., 2006) and syntax-based machine translation systems (Galley *et al*., 2004; Galley *et al.*, 2006).

* CLCLP, TIGP, Academia Sinica, Taipei, Taiwan
+ Department of Computer Science, NTHU, Hsinchu, Taiwan
E-mail: {u901571; jason.jschang}@gmail.com

[Received June 6, 2008; Revised May 12, 2009; Accepted May 15, 2009]

Since the pioneering work of Brown *et al*. (1988), a myriad of research projects have focused on the statistical translation model. These could be classified into two main categories: one paying little attention to the grammar of the languages (Vogel *et al*., 1996; Och & Ney, 2000; Toutanova *et al*., 2002) and the other explicitly utilizing languages' structural or syntactic information (Wu, 1997; Yamada & Knight, 2001; Cherry & Lin, 2003; Gildea, 2004; Zhang & Gildea, 2005). With an increasing number of more accurate syntactic analyzers (*e.g*., part-of-speech tagger and Stanford parser) being developed and in view of the deficiency in modeling grammatical aspects of languages facing IBM-like models, the latter has received increasing attention.

Recently, in order to incorporate languages' syntax, Yamada and Knight (2001) transformed source-language (SL) (*e.g*., English) parse trees into target-language (TL) (*e.g*., Japanese) strings, using operations of reordering, inserting, and translating on tree nodes. Instead of accepting monolingual (*i.e*., SL or TL) parse trees to do the transformation, Wu's ITG model (1997) first associates production rules (*e.g*., S→NP VP) commonly shared by two languages with (straight or inverted) word orientation and, based on these synchronous rules, constructs bilingual parse trees at run time. This data-oriented parsing methodology is reported to outperform tree-to-string model (*i.e*., (Yamada & Knight, 2001)) concerning word-level alignment (Zhang & Gildea, 2004).

Even though the promising ITG is proposed, Wu (1997) conducts a word-aligning experiment leveraging a special case of ITG, minimal bracketing transduction grammar (BTG), in which languages' grammars are assumed to be unavailable, constituent categories (*e.g*., NP and VP) are not differentiated (using only three symbols: one for lexical translation rules, another for straight binary production rules, the other for inverted), and the probabilities of the straight and inverted binary rules are all assigned constant. These imply that the choices of straight or inverted word orientations would be made *solely* based on the bonds of lexical translations rather than on the structural divergences of the involved languages and that the potential of the syntax-oriented ITG would not be fully explored.

More recently, Zhang and Gildea (2005) presented a lexicalized BTG model where orientation choices are also dependent on the head words of the structural constituents. They expect lexical pairs passed up from the bottom (*i.e.*, leaf nodes) of the bilingual parse tree will make BTG models more knowledgeable in determining straight/inverted word order. Nonetheless, they found that lexical information at the lower levels of trees is more deterministic in word orientations than that at the higher levels.

To explore the power of ITG a little more (and inspired by Zhang *et al*. (2006), who suggest that binarized rules improve both speed and accuracy of a syntax-based machine translation system), in this paper, we describe a version of ITG model where the binary grammatical rules (*e.g*., S→NP VP) of the source language (*e.g*., English) are used as the

skeleton of our synchronous rules. Since the rules are biased toward the syntactic labels of the source language, our model is referred to as *bITG* model, short for biased ITG model. In our model, based on word-aligned sentence pairs, binary SL CFG rules are automatically annotated with the target language's word orientations and the associated orientation probabilities are automatically computed via Maximum Likelihood Estimation (MLE).

For example, take the languages of English, Chinese, and Japanese. The higher probability of our binary bITG rule VP→[VP NP], where the square brackets denote the same ordering (*straight*) of the two right-hand-side constituents in both languages when expanding the left-hand-side symbol, indicates a similar VO construct exists in English (SVO language) and Chinese (SVO language). On the contrary, the different VO construct in English and Japanese (SOV language) is modeled through the high *inverted* probability of the binary bITG rule VP→<VP NP> where the pointed brackets denote that we expand the left-hand-side symbol into two right-hand-side symbols in reverse orientation in two languages. Notice that these two bITG rules originate from *the same* binary CFG rule (VP→VP NP) of the source language, English, only with *different* ordering tendencies on the TL (*i.e.*, Chinese or Japanese) end.

In addition, we leverage IBM-style fertility probabilities of words to accommodate many-to-one or one-to-many word alignment links. In other words, in our model, many contiguous words in the source can be aligned to one word in the target and vice-versa. Originally, Wu's BTG model (1997) only allowed for a maximum of one-to-one word correspondences, which may affect the performance on word alignments and the accuracy of the bilingual parse trees. This one-to-one mapping restriction is especially not suitable for a language pair involving a language without clear word delimiters since the tokenization (or segmentation) of sentences of that language (*e.g.*, Chinese) prior to word alignment is independent of words of another (*e.g.*, English), resulting in tokens being under- or over-segmented for the corresponding words and, subsequently, abundant many-to-one/one-to-many word alignments.

The paper is organized as follows. Sections 2 and 3 describe our model in detail. Section 4 shows empirical results. Discussions are made before the conclusion in Section 6.

2. Method

In this section, we begin with an example of how bITG rules and fertilities of words are utilized to assist in word-aligning sentence pairs. Thereafter, a more formal description of our model will be discussed.

2.1 An Example

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

English POS tags: DT NNS MD VB TO VB DT JJ NN IN PRP$ NN

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

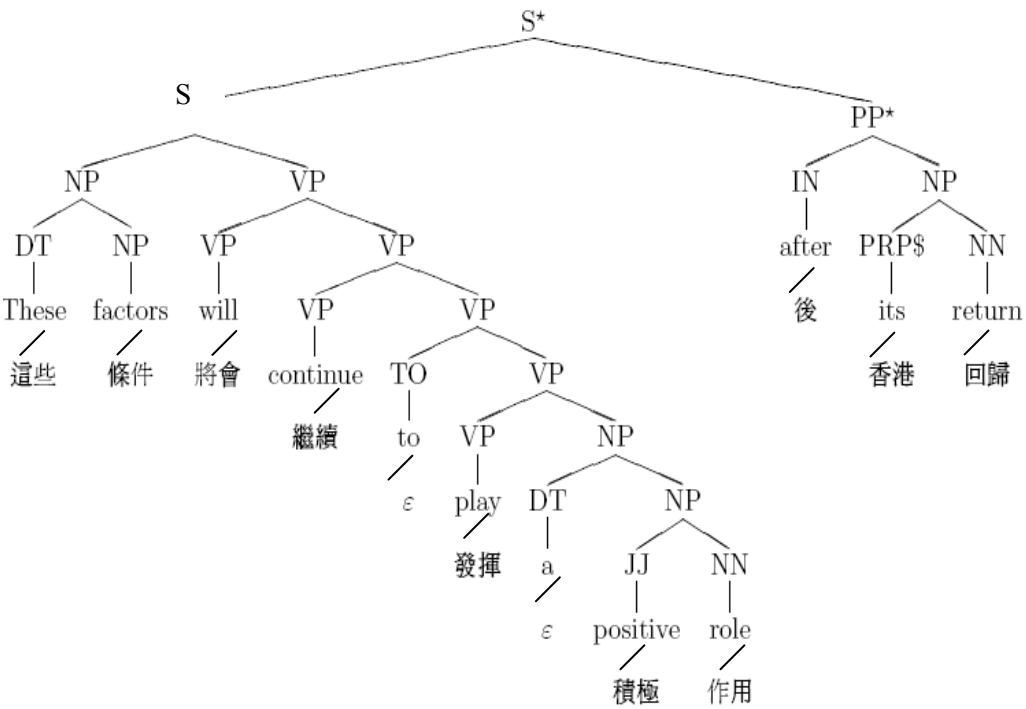

Figure 1. An example sentence pair and its bilingual parse tree

Once a sentence pair and the part-of-speech (POS) information of the SL sentence are fed into our model, it synchronously parses the sentence pair using unary lexical translation rules (*e.g.*, JJ→positive/積極 where / denotes word correspondence in two languages) and binary SL CFG rules attached with orientation preferences in the target language (*e.g.*, VP→[VP NP]). Also, the leaves of the bilingual parse tree are the word alignment results for this sentence pair.

During bilingual parsing, the model assigns probabilities to substring pairs of the bitext after each of them is associated with possible syntactic labels on the source side. For example, take the sentence pair and its parse in Figure 1, where spaces in the Chinese sentence are used to distinguish the boundaries of segments, ε stands for NULL, and * denotes the inverted orientation of the node's two children on the target. The substring pair (positive role, 積極 作用) associated with linguistic symbol NP will be assigned a probability. In this particular parse, the probability is the product of probabilities of the straight binary bITG rule, NP→[JJ NN],

and the lexical rules of bITG, JJ→positive/積極, and NN→role/作用. In our model, the higher probability of rule NP→[JJ NN] than the probability of the corresponding inverted rule NP→<JJ NN> does not merely instruct the model to align the two right-hand-side counterparts (*i.e.*, JJ and NN) of two languages in a straight fashion *more*, but also implies English and Chinese exhibit similar word-order regularity regarding the syntactic constituents.

On the other hand, in the example sentence pair, 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. Inverted rules (*e.g.*, S→<S PP>) are designed to capture such systematic differences in the languages' grammars.

What is more, since only monolingual information is exploited to segment Chinese sentences, it is likely that the word alignments will *not* be constrained to one-to-one, one-to-zero, and zero-to-one mappings. For instance, *香港* is often segmented as *a* word in Chinese but needs to be aligned to two words (*Hong* and *Kong*) in English, a case of two-to-one mapping. Therefore, we incorporate notion of fertility into our model.

As for the example of "*Hong Kong*" aligned to "*香港*", three possible word-aligning scenarios concerning fertility will be considered at runtime parsing: zero fertility of *Hong* and singular fertilities of *Kong* and *香港* where *Hong* is aligned to NULL but *Kong* is aligned to *香港*; zero fertility of *Kong* and singular fertilities of *Hong* and *香港* where *Kong* is aligned to NULL but *Hong* is aligned to *香港*; singular fertilities of *Hong* and *Kong* and dual fertility of *香港* where both *Hong* and *Kong* are aligned to *香港*.

Taking into account the probabilities of lexical translations, binary grammatical rewrite rules, and fertilities of words, our model manages to find a better parse tree that applies more appropriate synchronous rules to match the structural divergences and more suitable lexical mapping relations (one-to-one, one-to-two, *et al.*) in two languages. Better parses are more likely to yield better word alignment results.

We actually estimate the probabilities of bITG rules, consisting of unary lexical translation rules and binary SL CFG rules with word orientation on the TL, and those of the fertilities of words from a parallel corpus and an SL CFG. We will discuss the training algorithm in more detail in Section 3.

2.2 Formal Description

We now formally describe our statistical translation model. To be comparable to previous work, 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*.

Given $(e, f) = (e_1 \cdots e_m, f_1 \cdots f_n)$ and the POS tag sequence of e, τ, our model aims

to construct the most probable bilingual parse tree B_t^*, satisfying $\arg\max_{B_t}\{\Pr(B_t|e, f, \tau)\}$, with the by-product of word-level correspondences. Intuitively, the probability of a bilingual parse tree B_t provided with *e*, *f*, and τ is modeled as the product of probabilities associated with grammatical rewrite rules and lexical information:

$$\Pr(B_t|e, f, \tau) = \Pr(\mathbf{D}|e, f, \tau) \times \Pr(\mathbf{A}|e, f, \tau) \quad (1)$$

where, by inspecting the parse tree B_t, **D,** and **A** represent the set of its production rules with syntactic labels on the right hand side (*e.g.*, NP→JJ NN) and the set of rules with word alignments on the right (*e.g.*, JJ→positive/積極), respectively.

For simplicity, we use α_k to denote internal nodes (NP, JJ, *etc*) of the tree B_t, whereas we use β_k to denote leaf nodes (*e.g.*, these/這些, positive/積極). Tree nodes in B_t can be divided into three groups according to the number of children they are connected to: the first, denoted by set $\mathbf{N}_2$, consists of nodes with two children; the second, denoted by set $\mathbf{N}_1$, is made up of nodes with one child; the last, denoted by set $\mathbf{N}_0$, comprises nodes without a child. For notation convenience, each $\alpha_k \in \mathbf{N}_2$ has two children represented by α_{2k} and α_{2k+1}, and each $\alpha_k \in \mathbf{N}_1$ has one child β_k.

In our model, the probability of constructing B_t is the product of the probabilities of two sources: the first estimating the probabilities of the applied binary bITG rules; the second estimating those of the unary lexical translation rules and the fertilities of words in the tree. Assuming each applied rule is independent of one another, we rewrite the grammatical-related term in Equation (1) as

$$\Pr(\mathbf{D}|e, f, \tau) \cong \prod_{\alpha_k \in \mathbf{N}_2} \mathrm{P}^{\lambda_1}(\alpha_k \rightarrow [\![\alpha_{2k}\ \alpha_{2k+1}]\!]) \quad (2)$$

where $[\![\ ]\!]$ can be straight $[\ ]$ or inverted $\langle\ \rangle$. On the other hand, the lexical-related term in Equation (1) is decomposed into three factors, as shown in Equation (3): one for the product of probabilities of lexical translation rules given τ, another for the product of fertility probabilities of words in *e,* and the other for the product of fertility probabilities of words in *f*.

$$\Pr(\mathbf{A}|e, f, \tau) \cong \prod_{\alpha_k \in \mathbf{N}_1} \mathrm{P}^{\lambda_2}(\alpha_k \rightarrow \beta_k|\tau) \times \prod_{i=1}^{m} \mathrm{P}^{\lambda_2}(\Phi = \phi_{e_i}) \times \prod_{j=1}^{n} \mathrm{P}^{\lambda_2}(\Phi = \phi_{f_j}) \quad (3)$$

In Equation (3), Φ is the random variable for fertilities of words, and ϕ_{e_i} and ϕ_{f_j} denote fertilities of e_i and f_j, respectively. From Equations (1) to (3), we estimate the probability of a parse tree via

$$\Pr(B_t|e,f,\tau) \cong \prod_{\alpha_k \in \mathbf{N_2}} \mathrm{P}^{\lambda_1}\left(\alpha_k \to [\![\alpha_{2k}\ \alpha_{2k+1}]\!]\right) \times \prod_{\alpha_k \in \mathbf{N}_1} \mathrm{P}^{\lambda_2}\left(\alpha_k \to \beta_k|\tau\right) \times \prod_{i=1}^{m} \mathrm{P}^{\lambda_2}\left(\Phi_{e_i} = \phi_{e_i}\right) \times \prod_{j=1}^{n} \mathrm{P}^{\lambda_2}\left(\Phi_{f_j} = \phi_{f_j}\right) \quad (4)$$

in which the sum of the weight λ_1 and λ_2 is one.

2.3 Runtime Parsing

In this subsection, we depict a CYK-like parsing algorithm for obtaining the most likely bilingual parse tree given the sentence pair $(e,f)=(e_1\cdots e_m, f_1\cdots f_n)$, the pre-determined POS tag sequence, $(t_1,\cdots,t_m)$, of sentence *e*, and the grammar *G* in *E* (*i.e.*, SL grammar). Notice that our model is a data-driven one as is Wu (1997). In other words, it synchronously parses sentence pair via bITG rules *without* a monolingual (SL or TL) parse tree. Figure 2 shows the run-time parsing algorithm.

Parsing Algorithm

//Initial Step

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

(1) $\delta_{t_i,i-1,i,j-1,j} = \mathrm{P}^{\lambda_2}\left(t_i \to e_i/f_j\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{e_i}=1\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{f_j}=1\right)$

(2) For every $L \to t_i \in G$ in E

(3) $\delta_{L,i-1,i,j-1,j} = \mathrm{P}^{\lambda_2}\left(L \to e_i/f_j\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{e_i}=1\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{f_j}=1\right)$

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

(4) $\delta_{t_i,i-1,i,j,j} = \mathrm{P}^{\lambda_2}\left(t_i \to e_i/\varepsilon\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{e_i}=0\right)$

(5) For every $L \to t_i \in G$ in E

(6) $\delta_{L,i-1,i,j,j} = \mathrm{P}^{\lambda_2}\left(L \to e_i/\varepsilon\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{e_i}=0\right)$

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

(7) $\delta_{L,i,i,j-1,j} = \mathrm{P}^{\lambda_2}\left(L \to \varepsilon/f_j\right) \times \mathrm{P}^{\lambda_2}\left(\Phi_{f_j}=0\right)$

//Recurrent Step

For any possible (s,t,u,v) // $1 \le s,t \le m, 1 \le u,v \le n$

For any possible grammatical label *p*

If ($t \ge s$ and $v \ge u$) and not ($t = s$ and $v = u$)

(8)
$$\delta_{p,s,t,u,v} = \max_{\substack{q,r\in\text{syntax labels on } E \\ s\le s'\le t \\ u\le u'\le v}} \left\{ \begin{array}{l} \mathrm{P}^{\lambda_1}\left(p \to [q\ r]\right) \times \delta_{q,s,s',u,u'} \times \delta_{r,s',t,u',v}, \\ \mathrm{P}^{\lambda_1}\left(p \to \langle q\ r \rangle\right) \times \delta_{q,s,s',u',v} \times \delta_{r,s',t,u,u'} \end{array} \right\}$$

//for backtracking

$$\mathrm{b}_{p,s,t,u,v} = \arg\max_{\substack{q,r\in\text{syntax labels on } E \\ s\le s'\le t \\ u\le u'\le v}} \left\{ \begin{array}{l} \mathrm{P}^{\lambda_1}\left(p \to [q\ r]\right) \times \delta_{q,s,s',u,u'} \times \delta_{r,s',t,u',v}, \\ \mathrm{P}^{\lambda_1}\left(p \to \langle q\ r \rangle\right) \times \delta_{q,s,s',u',v} \times \delta_{r,s',t,u,u'} \end{array} \right\}$$

(9) Backtrack()

Figure 2. Run-time parsing.

During a parse of a sentence pair in our model, a table of $\delta_{p,s,t,u,v}$, the *best* probability for parsing substring pair $(e_{s+1}\cdots e_t, f_{u+1}\cdots f_v)$ attached with a syntactic symbol p on E side, is constructed.

In Step (1) of Figure 2, we compute the probability of a one-to-one word correspondence e_i/f_j with e_i's pre-determined POS tag t_i, according to the probability of the unary bITG rule $t_i \to e_i/f_j$ and the probabilities of fertilities of e_i and f_j (fertilities are 1s for one-to-one mapping). Since the POS tag t_i can be derived from some possible phrasal constituents in *G* (Step (2)) (*e.g.*, NN can be derived from NP), we also compute their associated probabilities (Step (3)). Similarly, in Steps (4) to (7), we calculate the probabilities of the one-to-zero and zero-to-one word correspondences limited to the scope of the sentence pair.

Afterwards, relying on the work done previously, word correspondences and parsing results of longer substring pairs would unveil themselves in a bottom-up manner. In Step (8), *s*' divides the substring $e_{s+1}\cdots e_t$, labeled as *p*, into two parts, $e_{s+1}\cdots e_{s'}$ and $e_{s'+1}\cdots e_t$, with *q* as a possible grammatical symbol of the first part and *r* as a possible symbol of the second, while *u*' divides the substring $f_{u+1}\cdots f_v$ into $f_{u+1}\cdots f_{u'}$ and $f_{u'+1}\cdots f_v$. As the substring $e_{s+1}\cdots e_{s'}$ can be aligned to $f_{u+1}\cdots f_{u'}$ or $f_{u'+1}\cdots f_v$, both straight and inverted orientation of the SL CFG rules "$p \to q\ r$" ought to be considered. Note that the computation in Step (8) does not properly deal with the cases of many-to-one or one-to-many word-level alignments. For many-to-one alignments, $\delta_{p,s,t,u-1,u}$ should further incorporate the parsing candidate:

$$\mathrm{P}^{\lambda_2}\left(\Phi_{f_u} = (t-s)\right) \times \max_{\substack{q,r\in\text{syntax} \\ \text{labels on } E}} \left\{ \mathrm{P}^{\lambda_1}\left(p \to [q\ r]\right) \times \frac{\delta_{q,s,s+1,u-1,u}}{\mathrm{P}^{\lambda_2}\left(\Phi_{f_u} = 1\right)} \times \frac{\delta_{r,s+1,t,u-1,u}}{\mathrm{P}^{\lambda_2}\left(\Phi_{f_u} = (t-s-1)\right)} \right\}$$

where $\delta_{r,s+1,t,u-1,u}$ needs to be constructed from many-to-one or one-to-one word mapping relation since words $e_{s+1}\cdots e_t$ are all aligned to f_u. A similar principal applies to one-to-many mapping (*i.e.*, the calculation of $\delta_{p,s-1,s,u,v}$).

Finally, using the standard CYK backtracking technique, we can find the most probable bilingual parse tree of the sentence pair with word alignment results. The integration of fertilities of words into the model aims to improve the parsing and the word-aligning quality.

2.4 Pruning

Although the complexity of the described algorithm is polynomial-time (proportion to m^3n^3), the execution time grows rapidly with the increase in the variety of syntactic labels, from three structural labels (Wu, 1997) to the grammatical categories of the source language's syntax in our model. As a result, pruning techniques are essential to reduce the time spent on parsing.

We adopt pruning in the following two manners. The first pruning technique is, for a given SL substring $e_{s+1}\cdots e_t$ and a given TL substring's length, to only keep parse trees whose probabilities fall within the *best* $N\times\sigma$, where N is the number of possible parses for a SL substring $e_{s+1}\cdots e_t$ and a length of the TL substring, and σ is a real number between 0 and 1. In other words, we remove inferior parse trees that are not in the set of the best $N\times\sigma$ ones. Since *N* varies from case to case (depending on the SL substring and the length of TL substring), only the more probable trees within the ratio (*i.e.*, σ) of *N* will remain.

The second pruning technique is related to the ratio of the length of the SL and TL substring. $\delta_{p,s,t,u,v}$ will not be calculated if $\frac{t-s}{v-u}$ 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.

By applying the aforementioned pruning techniques, the time spent on parsing each sentence pair can be reduced by *more than* half. Empirically, pruning unlikely parses has little affect on the word alignment quality but reduces computational overhead significantly.

3. Probability Estimation

In this section, we describe how to estimate the probabilities of our unary bITG rules (*e.g.*, JJ→positive/積極) and binary bITG rules (*e.g.*, VP→[VP NP]) which denote the association of bilingual lexical words and model the structural divergences of the two languages, respectively. Figure 3 shows the probabilistic estimation procedure.

(1) $\mathbf{H} = \mathbf{WA}$

(2) For $\left(r, e_{i_1}^{i_2}, f_{j_1}^{j_2}, L, rhs, rel\right) \in \mathbf{H}, \left(r, e_{i_1'}^{i_2'}, f_{j_1'}^{j_2'}, L', rhs', rel'\right) \in \mathbf{H}$ have not been considered

(3) If $\left(i_2 = i_1' - 1\right)$

(4) For every $L'' \rightarrow L\ L' \in G$ in E

(5) If $\left(j_2 + 1 \le j_1' \le j_2 + \delta\right)$

(6) $\mathbf{H} = \mathbf{H} \cup \left\{\left(r, e_{i_1}^{i_2'}, f_{j_1}^{j_2'}, L'', L\ L', Straight\right)\right\}$

(7) If $\left(j_2' + 1 \le j_1 \le j_2' + \delta\right)$

(8) $\mathbf{H} = \mathbf{H} \cup \left\{\left(r, e_{i_1}^{i_2'}, f_{j_1'}^{j_2}, L'', L\ L', Inverted\right)\right\}$

(9) similar strategy applies when $i_2' = i_1 - 1$

Incorporate words aligned to null, each of which is denoted using 6-tuple representation, in both languages into **H**.

For $\left(r, e_{i_1}^{i_2}, f_{j_1}^{j_2}, L, rhs, rel\right) \in \mathbf{H}$

If (rhs is not a lexical pair)

(10) $$\mathrm{P}\left(L \rightarrow \left[R_1\ R_2\right]\right) = \frac{\mathrm{count}\left(\left(*,*,*,L,R_1\ R_2, Straight\right); \mathbf{H}\right)}{\left|\mathbf{H}\right|}$$

(11) $$\mathrm{P}\left(L \rightarrow \left\langle R_1\ R_2\right\rangle\right) = \frac{\mathrm{count}\left(\left(*,*,*,L,R_1\ R_2, Inverted\right); \mathbf{H}\right)}{\left|\mathbf{H}\right|}$$

Else

(12) $$\mathrm{P}\left(L \rightarrow e_i / f_j\right) = \frac{\mathrm{count}\left(\left(*,*,*,L,e_i / f_j, *\right); \mathbf{H}\right)}{\left|\mathbf{H}\right|}$$

(13) Based on **WA** and **C**, estimate $\mathrm{P}\left(\Phi_{e_i}\right)$ and $\mathrm{P}\left(\Phi_{f_j}\right)$ via MLE

Figure 3. The procedure of probabilistic estimation.

In Step (1) of our training procedure, an existing word-aligning strategy or tool (*e.g.*, GIZA++) is employed to obtain the word alignments (*i.e.*, **WA**) of a parallel corpus **C**. **WA** comprises elements of the form $\left(r, e_{i_1}^{i_2}, f_{j_1}^{j_2}, L, rhs, rel\right)$, which represents that the 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 production rule leading to the bilingual structure and has *rel* (either *straight* or *inverted*) as the cross-language word-order relation of the constituents of *rhs*. *rhs* denotes either a sequence of syntactic labels or a terminating bilingual word pair. Following this format, the example parses of (positive,積極)$_{JJ}$ and (after its return,香港 回歸 後)$_{PP}$ in Figure 1 would be denoted by the 6-tuple (193, e_8^8, f_9^9 ,JJ,positive/積極,don't_care) and (193, e_{10}^{12}, f_1^3 ,PP,IN NP,*Inverted*) respectively, where 193 is the record number of this sentence pair.

Then, we recursively select two sections of a sentence pair, which have not yet been

paired up, from **H** (Step (2))**.** If the SL substring of the first section (*i.e.*, $e_{i_1}^{i_2}$) is adjacent to that of the second (*i.e.*, $e_{i_1'}^{i_2'}$) on the right (Step (3)), based on word alignment result (Step (5) and Step(7)), a new straight-ordered (Step (6)) or inverted-ordered (Step (8)) section representing these two will be added into **H**. Specifically, once the SL substrings are related to some possible binary SL CFG rules, the right-hand-side constituents of these rules will be associated with an orientation on the TL end based on word alignment links. Since our model is a synchronous bilingual parsing one, *without* a monolingual parse tree, it enumerates all possible syntactic symbols to derive L and L' in Step (4). Note that, in Steps (5) and (7), δ, a small positive integer, is utilized to tolerate aligning errors introduced by the automatic word aligner or explicitness issue[1] during translation from one language to another, when determining cross-language straight/inverted word order phenomenon.

From Step (10) to Step (12), in which |**W**| stands for the number of entries in set **W** and count(*p*;**Q**) for the frequency of *p* in set **Q**, we estimate probabilities of bITG rules via Maximum Likelihood Estimation. In our model, the probabilities of lexical translation rules (*e.g.*, JJ→positive/積極) and binary bITG rules (*e.g.*, VP→[VP NP]) are estimated from the same source (*i.e.*, **H**). Alternative probabilistic estimation of these two kinds of rules can be adopted. For example, the probabilities of lexical translation rules can be derived from pure word alignment set **WA** while those of binary bITG rules can be derived from set **H** without word-level alignment links. We employ the former estimation approach and, in experiments, it yields satisfying results (see Section 4), suggesting word-order tendencies of the two languages are properly modeled.

Finally, fertility probabilities related to words in both languages are also calculated (Step (13)).

4. Experiments

In experiments, we trained our model on a large English-Chinese parallel corpus. We examined word alignments produced by our bITG 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 parallel corpus.

4.1 Training Proposed Model

We used the news portion of Hong Kong Parallel Text[2] (HKPT) distributed by Linguistic Data Consortium as our sentence-aligned corpus **C,** which consisted of 739,919 English-Chinese

[1] Some translations may be omitted for conciseness, or some of the function words in one language may have no counterparts in another.

[2] LDC2004T08

sentence pairs. The average length was 24.4 words for English and 21.5 words for Chinese.

In our model, English sentences were considered to be the source while Chinese sentences were the target. SL sentences were POS tagged and TL sentences were segmented prior to word alignment. During training (as described in Section 3), we employed a GIZA++ run with default settings to obtain the word alignment set **WA** and our binary SL CFG *G* was based upon PTB section 23[3] production rules distributed by Andrew B. Clegg.

4.2 Evaluation

To evaluate our statistical translation model, 114 sentence pairs were chosen randomly from the news portion of HKPT as our testing data set. For the sake of execution time, we only selected sentence pairs whose SL and TL length did not exceed 15. Sentence pairs satisfying such a length constraint covered approximately 40% of the sentence pairs in the news portion of HKPT and were expected to be better word aligned via GIZA++.

We examined the word-aligning performance using the metrics of alignment error rate (AER) proposed by Och and Ney (2000), in which the quality of a word alignment result **A** produced by an automatic system is evaluated by:

$$precision = \frac{|\mathbf{A} \cap \mathbf{P}|}{|\mathbf{A}|}, \quad recall = \frac{|\mathbf{A} \cap \mathbf{S}|}{|\mathbf{S}|} \text{ and } \quad AER(\mathbf{S}, \mathbf{P}; \mathbf{A}) = 1 - \frac{|\mathbf{A} \cap \mathbf{S}| + |\mathbf{A} \cap \mathbf{P}|}{|\mathbf{A}| + |\mathbf{S}|}.$$

In AER, **S** (sure) denotes the set whose alignments are not ambiguous and **P** (possible) denotes the set consisting of alignments that might or might not exist $(\mathbf{S} \subseteq \mathbf{P})$. Thus, human annotations may contain many-to-one, one-to-many, or even many-to-many word alignments. Table 1 shows the experimental results of GIZA++, the BTG model (Wu, 1997), and our fertility-based SL-biased ITG model.

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
BTG	.844	.610	.290	.708
bITG w/o fertility	.866	.638	.263	.735
bITG w/ fertility	.878	**.692**	**.224**	**.774**

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

In this table[4], P, R, and F stand for precision, recall, and F-measure[5], respectively. The performance of the E-to-F alignments (E stands for English and F for Chinese), the F-to-E alignments, and the refined alignments (proposed by Och and Ney (2000)) from both E-to-F and F-to-E directions of GIZA++ are shown in first three rows, along with that of BTG, which also trained on the word-aligning 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 the BTG model that *does not* distinguish the constituent categories and makes the orientation choices merely on lexical evidence (without the information of languages' grammars), our model *without* fertility probability which allows for at most one-to-one alignment, as the BTG model does, achieved 9% reduction in the alignment error rate. This indicates that the binary SL CFG rules encoding with TL ordering preference in our model do capture the linguistic information of the languages such as word-order regularities or grammar and do impose more realistic and accurate reordering constraints on word alignment in the language pairs.

Furthermore, in comparison to the refined alignments of both word-aligning directions, our model *with* the concept of fertility (allowing for many-to-one/one-to-many links), which is quite similar to the refined approach accommodating many-to-many word mappings, increased the recall by 9% while maintaining high precision and achieved 14% alignment error reduction overall (increased F-measure by 5%).

As suggested by Table 1, it is safe to say that the proposed model yields more accurate bilingual parse trees, thus better word alignment quality, by introducing binary CFG rules of a language (*i.e.*, the source language) and fertility notation of IBM models into ITG model.

5. Discussion

In this section, we examine how the learnt similarities (*straight*) and differences (*inverted*) in word orders of two languages aid the word-aligning process of our model by means of the adjacency feature and cohesion constraint, mentioned in Cherry and Lin (2003). Subsequently, to evaluate the possibility of better machine translation quality by providing our model's output (*i.e.*, word correspondences), we adopt the recently-proposed metric, consistent phrase error rate (CPER) by Ayan and Dorr (2006).

---

[4] $\frac{|\mathbf{S}|}{|\mathbf{P}|}$ is 85.56% in human-annotated test data.

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

5.1 *Straight/Inverted* Orientation

Table 2 shows the accuracy of adjacent alignments made by our model, and the accuracy achieved by the refined approach is shown for comparison. If compared against the gold standard in the sure set (*i.e.*, **S** in Section 4), our model with bITG rules relatively increased the accuracy by more than 3%, suggesting the similar (or straight) word orientations of the binary syntactic constituents (*e.g.*, JJ and NN) in the languages are better captured in our model than in GIZA++. Note that alignments must have orders before an adjacency feature exists (see Cherry and Lin (2003)) in them. Therefore, an ordering, depending on the position of the English word in the sentence, was imposed to examine the feature.

Table 2. Examination of adjacent links

	Compared tosure links	Compared to possible links
Refined	.835	.869
bITG w/ fertility	**.863**	**.881**

Additionally, we examined whether the inverted binary bITG rules captured the diversities of the two grammars and helped to make correct crossing (or reverse) alignment links or not. For that purpose, we first acquired the dependency relations of the source (*i.e.*, English) sentences via a Stanford parser, and computed the percentage of links violating the cohesion constraint (see Cherry and Lin (2003)). The ratios of having crossing dependencies in the mapped Chinese dependency trees[6] are summarized in Table 3. As suggested by Table 3, our model reduced sixteen percent of the links violating the cohesion constraint (compared to the refined approach).

Table 3. Percentage of links violating cohesion constraint

	Percentage
Refined	.044
bITG w/ fertility	**.037**

The above statistics indicate that the probabilities related to straight and inverted word orders of bITG rules in our model not only impose a more suitable alignment constraint but properly model the systematic similarities and differences in two languages' grammars.

[6] Chinese dependency trees are mapped from English dependency trees based on word correspondences.

5.2 CPER

According to Ayan and Dorr (2006), the intrinsic evaluation metric of AER (Och and Ney, 2000) examines only the quality of word-level alignments and correlates poorly with the MT-community metric—BLEU score. As a result, we exploited consistent phrase error rate (CPER) to evaluate words alignments in the context of machine translation. CPER is reported to better correlate with translation quality (the smaller the CPER is, the better the translation quality) in that it evaluates phrase-level alignments and in that phrase-level alignments (bilingual phrase pairs) constitute the key essences of a MT system.

In Ayan and Dorr (2006), precision (P), recall (R), and CPER are computed via:

$\mathrm{P} = \frac{|P_A \cap P_G|}{|P_A|}$, $\mathrm{R} = \frac{|P_A \cap P_G|}{|P_G|}$, and $\mathrm{CPER} = 1 - \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}}$ where P_A and P_G stand for two sets of phrases generated by an automatic alignment *A* and manual alignment *G*, respectively. In Table 4, the proposed fertility-based source-language-based ITG model yielded the lowest CPER. This indicates that MT systems, accepting our word alignment output, are more likely to lead to better translation performance.

Table 4. Reports on CPER

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

6. Conclusion and Future Work

To combine the strengths of the competing models, a thought-provoking fusion of IBM-style fertility with syntax-based ITG model is described. In our model, the orientation probabilities of the binary SL-based ITG rules are automatically estimated based on a word-aligned parallel corpus and are devised to better capture structural divergences of the involved languages. The proposed bITG model with fertility reduces AER by 14% and 23%, and reduces CPER by 9% and 13% compared to GIZA++ and Wu's BTG (1997), respectively. Lower CPER suggests MT systems chained after our statistical translation model are likely to yield better translation quality. In this paper, the performance of ITG models trained on large-scale bitexts is shown

for the first time with quite encouraging results.

As for future work, we would like to explore methods (*e.g*. (Brown, 1992)) for partitioning long sentences into shorter ones so that the time spent on bilingual parsing in our model can be reduced. We also like to see whether word-aligning quality can be further improved if our bITG rules are lexicalized, especially when lexical contents play an important role in determining word orders of the languages.

References

Ayan, N. F. & Dorr, B. J. (2006). Going beyond AER: an extensive analysis of word alignments and their impact on MT. In *Proceedings of ACL-2006*, 9-16.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., Mercer, R. L., & Mohanty, S. (1992). Dividing and conquering long sentence in a translation system. In *Proceedings of the Workshop on Speech and Natural Language*, 267-271.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2), 263-311.

Cherry, C. & Lin, D. (2003). A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 88-95.

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

Clegg, A. B. & Shepherd, A. (2005). Evaluating and integrating Treebank parsers on a biomedical corpus. In *Association for Computational Linguistics Workshop on software 2005*.

Galley, M., Hopkins, M., Knight, K., & Marcu, D. (2004). What's in a translation rule? In *Proceedings of HLT/NAACL-2004*, 273-280.

Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W. et al. (2006). Scable inference and training of context-rich syntactic translation models. In *Proceedings of the 44th Annual Conference of the Association for Computational Linguistics*, 961-968.

Gildea, D. (2004). Dependencies vs. constituents for tree-based alignment. In *Proceedings of the EMNLP*, 214-221.

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

Och, F. J. & Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Conference of ACL-2000*, 440-447.

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

Toutanova, K., Ilhan, H. T., & Manning, C. D. (2002). Extentions to HMM-based statistical word alignment models. In P*roceedings of the Conference on Empirical Methods in Natural Processing Language*, 87-94.

Vogel, S., Ney, H., & Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, 836-841.

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

Yamada, K. & Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Annual Conference of ACL-2001*, 523-530.

Zens, R. & Ney, H. (2003). A comparative study on reordering constraints in statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 144-151.

Zhang, H. & Gildea, D. (2004). Syntax-based alignment: supervised or unsupervised? In P*roceedings of the 20th International Conference on Computational Linguistics,* 418-424.

Zhang, H. & Gildea, D. (2005). Stochastic lexicalized inversion transduction grammar for alignment. In *Proceedings of the 43rd Annual Meeting of the ACL*, 475-482.

Zhang, H., Huang, L., Gildea, D., & Knight, K. (2006). Synchronous binarization for machine translation. In *Proceedings of the NAACL-HLT*, 256-263.