# Unsupervised Word Alignment by Agreement Under ITG Constraint

**Hidetaka Kamigaito**[1] **Akihiro Tamura**[2]
kamigaito@lr.pi.titech.ac.jp akihiro.tamura@nict.go.jp
**Hiroya Takamura**[1] **Manabu Okumura**[1] **Eiichiro Sumita**[2]
takamura@pi.titech.ac.jp oku@pi.titech.ac.jp eiichiro.sumita@nict.go.jp
[1]Tokyo Institute of Technology
[2]National Institute of Information and Communication Technology

## Abstract

We propose a novel unsupervised word alignment method that uses a constraint based on Inversion Transduction Grammar (ITG) parse trees to jointly unify two directional models. Previous agreement methods are not helpful for locating alignments with long distances because they do not use any syntactic structures. In contrast, the proposed method symmetrizes alignments in consideration of their structural coherence by using the ITG constraint softly in the posterior regularization framework (Ganchev et al., 2010). The ITG constraint is also compatible with word alignments that are not covered by ITG parse trees. Hence, the proposed method is robust to ITG parse errors compared to other alignment methods that directly use an ITG model. Compared to the HMM (Vogel et al., 1996), IBM Model 4 (Brown et al., 1993), and the baseline agreement method (Ganchev et al., 2010), the experimental results show that the proposed method significantly improves alignment performance regarding the Japanese-English KFTT and BTEC corpus, and in translation evaluation, the proposed method shows comparable or statistical significantly better performance on the Japanese-English KFTT and IWSLT 2007 corpus.

## 1 Introduction

Word alignment is an important component of statistical machine translation (SMT) systems such as phrase-based SMT (Koehn et al., 2003) and hierarchical phrase-based SMT (Chiang, 2007). In addition, word alignment is utilized for multi-lingual tasks other than SMT, such as bilingual lexicon extraction (Liu et al., 2013). The most conventional approaches to word alignment are the IBM models (Brown et al., 1993) and the HMM model (Vogel et al., 1996), which align each source word to a single target word (i.e., directional models). In these models, bidirectional word alignments are traditionally induced by combining the Viterbi alignments in each direction using heuristics (Och and Ney, 2003). Matusov et al. (2004) exploited a symmetrized posterior probability for bidirectional word alignments. In these methods, each directional model is independently trained.

Previous researches have improved bidirectional word alignments by jointly training two directional models to agree with each other (Liang et al., 2006; Graça et al., 2008; Ganchev et al., 2010). Such a constraint on the agreement in a training phase is one of the most effective approaches to word alignment. However, none of the previous agreement constraints have taken into account syntactic structures. Therefore, they have difficulty recovering the alignments with long distances, which frequently occur, especially in grammatically different language pairs.

Some unsupervised word alignment models such as DeNero and Klein (2007) and Kondo et al. (2013), have been based on syntactic structures. In particular, it has been proven that Inversion Transduction Grammar (ITG) (Wu, 1997), which captures structural coherence between parallel sentences, helps in word alignment (Zhang and Gildea, 2004; Zhang and Gildea, 2005). However, ITG has not been introduced into an agreement constraint so far.

*Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2004,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

We propose an alignment method that uses an ITG constraint to encourage agreement between two directional models in consideration of their structural coherence. Our ITG constraint is based on the Viterbi alignment decided by a bracketing ITG parse tree, and used as a soft constraint in the posterior regularization framework (Ganchev et al., 2010). In addition, our ITG constraint works also on word alignments that are not covered by ITG parse trees, as a standard symmetric constraint. Hence, the proposed method is robust to ITG parse errors compared to an alignment method that uses an ITG directly in model training (e.g., Zhang and Gildea (2004, 2005)).

Word alignment evaluations show that the proposed method achieves significant gains in F-measure and alignment error rate (AER) on the KFTT (Neubig, 2011) and the BTEC Japanese-English (Ja-En) corpus (Takezawa et al., 2002). Machine translation evaluations show that our constraint significantly outperforms or is comparable to the baseline symmetric constraint (Ganchev et al., 2010) in BLEU on the KFTT Ja-En and IWSLT 2007 Ja-En corpus (Fordyce, 2007).

# 2 ITG Constraint in the Posterior Regularization Framework

## 2.1 Overview

The proposed method introduces an ITG constraint into the posterior regularization framework (Ganchev et al., 2010) in model training. The proposed model is trained as follows, where agreement constraints are imposed in the E-step of the EM algorithm[1]:

**E-step**:

1. Calculate a source-to-target posterior probability $\overrightarrow{p_\theta}(\boldsymbol{z}|\boldsymbol{x})$ and a target-to-source posterior probability $\overleftarrow{p_\theta}(\boldsymbol{z}|\boldsymbol{x})$ for each bilingual sentence $\boldsymbol{x} = \{\boldsymbol{f}, \boldsymbol{e}\}$ under the current model parameters $\boldsymbol{\theta}$, where $\boldsymbol{z}$ denotes an alignment in a sentence pair $\boldsymbol{x}$. In particular, $z_{i,j}$=1, if $f_i$ is aligned to $e_j$ (otherwise $z_{i,j}$=0).

2. Repeat the following steps for all sentence pairs in the training data.

(a) Find the Viterbi alignment $\boldsymbol{z}^*$ through ITG parsing (see Section 2.2). Here, $z^*_{i,j}$=1, if $f_i$ is aligned to $e_j$ (otherwise $z^*_{i,j}$=0).

(b) Symmetrize $\overrightarrow{p_\theta}(\boldsymbol{z}|\boldsymbol{x})$ and $\overleftarrow{p_\theta}(\boldsymbol{z}|\boldsymbol{x})$ under the constraint of $\boldsymbol{z}^*$ (see Section 2.3 and 2.4).

**M-step**:

1. Estimate all parameters $\boldsymbol{\theta}$ based on the symmetrized posterior probabilities $\overrightarrow{q_{\boldsymbol{\lambda}}}(\boldsymbol{z}|\boldsymbol{x})$ and $\overleftarrow{q_{\boldsymbol{\lambda}}}(\boldsymbol{z}|\boldsymbol{x})$ (see Section 2.3 and 2.4).

## 2.2 ITG Parsing

In this section, we present our ITG parsing method, which uses bracketing ITG (Wu, 1997). The rules of the bracketing ITG are as follows: $A \rightarrow \langle Y/Z \rangle$, $A \rightarrow [Y/Z]$, $A \rightarrow f_i/e_j$, $A \rightarrow f_i/\epsilon$, and $A \rightarrow \epsilon/e_j$, where $A$, $Y$, and $Z$ are non-terminal symbols, $f_i$ and $e_j$ are terminal strings, $\epsilon$ is a null symbol, $\langle\rangle$ denotes the inversion of two phrase positions, and [] denotes the reversion of two phrase positions.

In general, a bracketing ITG has $O(|\boldsymbol{f}|^3|\boldsymbol{e}|^3)$ time complexity for parsing a sentence pair $\{\boldsymbol{f}, \boldsymbol{e}\}$, where $|\boldsymbol{f}|$ and $|\boldsymbol{e}|$ are the lengths of $\boldsymbol{f}$ and $\boldsymbol{e}$. For efficient ITG parsing, we use the two-step parsing approach (Xiao et al., 2012), which has been proposed to induce Synchronous Context Free Grammar (SCFG) using n-best pruning[2] with time complexity $O(|\boldsymbol{f}|^3)$. Because ITG is a kind of SCFG, this method can be adopted for our ITG parsing. Our two-step parsing first parses a bilingual sentence in the bottom up manner, and then derives the Viterbi alignment $\boldsymbol{z}^*$ in the top down manner.

To parse a bilingual sentence $\boldsymbol{x} = \{\boldsymbol{f}, \boldsymbol{e}\}$, we define the probability for each ITG rule. The probability of a rule $A \rightarrow f_i/e_j$ is defined as:

$$P(A \rightarrow f_i/e_j) = \frac{\overrightarrow{p}_\theta(z_{i,j} = 1|\boldsymbol{x}) + \overleftarrow{p}_\theta(z_{i,j} = 1|\boldsymbol{x})}{2}.$$

We provide a constant value $p_{null}$[3] both to $P(A \rightarrow \epsilon/e_j)$ and $P(A \rightarrow f_i/\epsilon)$. To reduce computational cost, the probabilities of phrasal rules $P(A \rightarrow \langle Y/Z \rangle)$ and $P(A \rightarrow [Y/Z])$ are not trained, which are set to 0.5 following Saers et al. (2012). In addition to the probability of each ITG rule, we must provide a probability to an one-to-many alignment because the two step parsing approach must pre-compute probabilities for all one-to-many alignments in the first step. An one-to-many alignment

[1] Step 1 in the E and M steps can be performed in the same way as in Ganchev et al. (2010).

[2] We set n to 30 in our experiments.

[3] We set $p_{null}$ to $10^{-5}$.

can be decomposed to a rule $A \rightarrow f_i/e_j$ and some $A \rightarrow \epsilon/e_j$ rules under the ITG form. We select a set of rules with the highest probability for an one-to-many alignment using Viterbi algorithm, which has a complexity of $O(|\boldsymbol{e}|)$.

### 2.3 Previous Agreement Constraint

This section provides an overview of the previous agreement constraint proposed by Ganchev et al. (2010), which is our baseline. In the posterior regularization framework, source-to-target and target-to-source posterior probabilities $\overrightarrow{p}_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ and $\overleftarrow{p}_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ are replaced with $\overrightarrow{q}_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x})$ and $\overleftarrow{q}_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x})$, defined as follows:

$$\overrightarrow{q}_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x}) = \overrightarrow{p}_{\theta}(\boldsymbol{z}|\boldsymbol{x}) \cdot exp^{(-\boldsymbol{\lambda}\cdot\boldsymbol{\phi}^{\text{agree}}(\boldsymbol{x},\boldsymbol{z}))}/Z_{\overrightarrow{q}},$$
$$\overleftarrow{q}_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x}) = \overleftarrow{p}_{\theta}(\boldsymbol{z}|\boldsymbol{x}) \cdot exp^{(-\boldsymbol{\lambda}\cdot\boldsymbol{\phi}^{\text{agree}}(\boldsymbol{x},\boldsymbol{z}))}/Z_{\overleftarrow{q}},$$

where $Z_{\overrightarrow{q}}$ is a normalization term for $\sum_{\boldsymbol{z}} \overrightarrow{q}_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x}) = 1$ ($Z_{\overleftarrow{q}}$ is analogous) and $\boldsymbol{\lambda}$ is a vector of weight parameters that controls the balance between two directional posterior probabilities. Here, $\boldsymbol{\phi}^{\text{agree}}$ is a feature of agreement constraint, which assigns each alignment direction to a sign (i.e., +1 or -1). In particular, $\boldsymbol{\phi}^{\text{agree}}$ is defined as follows:

$$\phi_{i,j}^{\text{agree}}(\boldsymbol{x},\boldsymbol{z}) = \begin{cases} +1 & (\boldsymbol{z} \in \overleftarrow{\boldsymbol{Z}}) \wedge (z_{i,j}{=}1), \\ -1 & (\boldsymbol{z} \in \overrightarrow{\boldsymbol{Z}}) \wedge (z_{i,j}{=}1), \\ 0 & otherwise, \end{cases}$$

where $\overrightarrow{\boldsymbol{Z}}$ and $\overleftarrow{\boldsymbol{Z}}$ are sets of possible alignments generated by source-to-target and target-to-source alignment models, respectively. So that $\overrightarrow{q}_{\lambda_{i,j}}(z_{i,j}{=}1|\boldsymbol{x})$ and $\overleftarrow{q}_{\lambda_{i,j}}(z_{i,j}{=}1|\boldsymbol{x})$ become equal probabilities for each $i, j$ (i.e., $\overrightarrow{q}_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x})$ and $\overleftarrow{q}_{\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{x})$ are symmetrical), the agreement constraint is defined as follows:

$$\forall i, \forall j, \overrightarrow{q}_{\lambda_{i,j}}(z_{i,j}{=}1|\boldsymbol{x}) - \overleftarrow{q}_{\lambda_{i,j}}(z_{i,j}{=}1|\boldsymbol{x}) = 0. \quad (1)$$

To satisfy the constraint (1), each $\lambda_{i,j}$ is updated by a stochastic gradient descent in the E-step of EM algorithm.

### 2.4 Proposed ITG Constraint

This section presents the proposed ITG constraint based on the Viterbi alignment $\boldsymbol{z}^*$, which has previously been identified by the bracketing ITG parsing. The ITG constraint uses a feature $\boldsymbol{\phi}^{\text{ITG}}$ instead of $\boldsymbol{\phi}^{\text{agree}}$:

$$\phi_{i,j}^{\text{ITG}}(\boldsymbol{x},\boldsymbol{z}){=}\begin{cases} 0 & \overleftarrow{Y}(i,j){\wedge}(z_{i,j}^*{=}1){\wedge}(\delta_{i,j}(\boldsymbol{x},\boldsymbol{z}){<}0), \\ +1 & \overleftarrow{Y}(i,j){\wedge}(z_{i,j}^*{=}1){\wedge}(\delta_{i,j}(\boldsymbol{x},\boldsymbol{z}){>}0), \\ -1 & \overrightarrow{Y}(i,j){\wedge}(z_{i,j}^*{=}1){\wedge}(\delta_{i,j}(\boldsymbol{x},\boldsymbol{z}){<}0), \\ 0 & \overrightarrow{Y}(i,j){\wedge}(z_{i,j}^*{=}1){\wedge}(\delta_{i,j}(\boldsymbol{x},\boldsymbol{z}){>}0), \\ +1 & \overleftarrow{Y}(i,j){\wedge}(z_{i,j}^*{\neq}1), \\ -1 & \overrightarrow{Y}(i,j){\wedge}(z_{i,j}^*{\neq}1), \\ 0 & otherwise, \end{cases}$$

where $\overleftarrow{Y}(i,j) = (\boldsymbol{z} \in \overleftarrow{\boldsymbol{Z}}) \wedge (z_{i,j}{=}1)$, $\overrightarrow{Y}(i,j) = (\boldsymbol{z} \in \overrightarrow{\boldsymbol{Z}}) \wedge (z_{i,j}{=}1)$, and $\delta_{i,j}(\boldsymbol{x},\boldsymbol{z}) = \overrightarrow{p}_{\theta}(z_{i,j}{=}1|\boldsymbol{x}) - \overleftarrow{p}_{\theta}(z_{i,j}{=}1|\boldsymbol{x})$. Similarly to $\boldsymbol{\phi}^{\text{agree}}$, $\boldsymbol{\phi}^{\text{ITG}}$ is imposed on $\overrightarrow{q}_{\boldsymbol{\lambda}_{i,j}}(z_{i,j}{=}1|\boldsymbol{x})$ and $\overleftarrow{q}_{\boldsymbol{\lambda}_{i,j}}(z_{i,j}{=}1|\boldsymbol{x})$ under the constraint (1). If $z_{i,j}^* \neq 1$, our feature $\phi_{i,j}^{\text{ITG}}$ operates similarly to $\phi_{i,j}^{\text{agree}}$ according to the last three rules. If $z_{i,j}^*{=}1$, $\boldsymbol{\phi}^{\text{ITG}}$ adjusts probabilities of alignments $\overrightarrow{q}_{\lambda_{i,j}}(z_{i,j}{=}1|\boldsymbol{x})$ and $\overleftarrow{q}_{\lambda_{i,j}}(z_{i,j}{=}1|\boldsymbol{x})$ by increasing the lower probability without decreasing the higher probability according to the first four rules. For example, when $z_{i,j}^* = 1$ and $\overrightarrow{q}_{\lambda_{i,j}}(z_{i,j} = 1|\boldsymbol{x})$ is larger than $\overleftarrow{q}_{\lambda_{i,j}}(z_{i,j} = 1|\boldsymbol{x})$, $\overleftarrow{q}_{\lambda_{i,j}}(z_{i,j}{=}1|\boldsymbol{x})$ is increased until $\overleftarrow{q}_{\lambda_{i,j}}(z_{i,j}{=}1|\boldsymbol{x})$ equals $\overrightarrow{q}_{\lambda_{i,j}}(z_{i,j}{=}1|\boldsymbol{x})$ according to the second and fourth rules. When $z_{i,j}^*{=}1$ and $\overleftarrow{q}_{\lambda_{i,j}}(z_{i,j}{=}1|\boldsymbol{x})$ is larger than $\overrightarrow{q}_{\lambda_{i,j}}(z_{i,j}{=}1|\boldsymbol{x})$, $\overrightarrow{q}_{\lambda_{i,j}}(z_{i,j}{=}1|\boldsymbol{x})$ is increased until $\overleftarrow{q}_{\lambda_{i,j}}(z_{i,j} = 1|\boldsymbol{x})$ equals $\overrightarrow{q}_{\lambda_{i,j}}(z_{i,j} = 1|\boldsymbol{x})$ according to the first and third rules. As a result, probabilities of word alignments in $\boldsymbol{z}^*$ tend to be higher than those of the other alignments.

| Task | Corpus | Train | Dev | Test |
|---|---|---|---|---|
| Word Alignment | Hansard | 1.13M | 37 | 447 |
| | KFTT | 330k | 653 | 582 |
| | BTEC | 10k | 0 | 10k |
| Machine Translation | KFTT | 330k | 1.17k | 1.16k |
| | IWSLT2007 | 40k | 2.5k | 489 |

**Table 1:** The numbers of parallel sentences for each data set.

## 3 Evaluation

We compared our proposed ITG constraint (*itg*) with the baseline agreement constraint (Ganchev et al., 2010) (*sym*) on word alignment and machine translation tasks. In word alignment evaluations, we used the French-English (Fr-En) Hansard Corpus (Mihalcea and Pedersen, 2003), Ja-En KFTT[4] (Neubig,

[4] We used the cleaned dataset distributed on the KFTT official web site (http://www.phontron.com/kftt/index.html).

| | Hansard Fr-En | | KFTT Ja-En | | BTEC Ja-En | |
|---|---|---|---|---|---|---|
| Method | F-measure | AER | F-measure | AER | F-measure | AER |
| HMM+*none* | 0.7900 | 0.0646 | 0.4623 | 0.5377 | 0.4425 | 0.5575 |
| HMM+*sym* | **0.7923** | **0.0597** | 0.4678 | 0.5322 | 0.4534 | 0.5466 |
| HMM+*itg* | 0.7869 | 0.0629 | 0.4690 | 0.5310 | 0.4499 | 0.5501 |
| IBM Model 4+*none* | 0.7780 | 0.0775 | 0.5379 | 0.4621 | 0.4454 | 0.5546 |
| IBM Model 4+*sym* | 0.7800 | 0.0693 | 0.5545 | 0.4455 | 0.4761 | 0.5239 |
| IBM Model 4+*itg* | 0.7791 | 0.0710 | **0.5613** | **0.4387** | **0.4809** | **0.5191** |

**Table 2:** Word alignment performance.

| Method | KFTT Ja-En | IWSLT2007 Ja-En |
|---|---|---|
| HMM+*none* | 18.9 | 46.4 |
| HMM+*sym* | 18.9 | 46.3 |
| HMM+*itg* | 19.2 | **47.0** |
| IBM Model 4+*none* | 18.8 | 46.7$^\dagger$ |
| IBM Model 4+*sym* | 19.3$^\dagger$ | 45.9 |
| IBM Model 4+*itg* | **19.4** | 46.7 |

**Table 3:** Machine translation performance.

2011), and Ja-En BTEC Corpus (Takezawa et al., 2002). We used the first 10K sentence pairs in the training data for the IWSLT 2007 translation task, which were manually annotated with word alignment (Chooi-Ling et al., 2010), as the BTEC Corpus. In translation evaluations, we used the KFTT and Ja-En IWSLT 2007 translation tasks[5].

Table 1 shows each corpus size. In each training data set, all words were lowercased and sentences with over 80 words on either side were removed.

### 3.1 Word Alignment Evaluation

We measured the performance of word alignment with AER and F-measure (Och and Ney, 2003). We used only sure alignments for calculating F-measure (Fraser and Marcu, 2007)[6]. We introduced *itg* and *sym* into the HMM and IBM Model 4. Training is bootstrapped from IBM Model 1, followed by HMM and IBM Model 4. All models were trained with five consecutive iterations. In the many-to-many alignment extraction, we used the filtering method (Matusov et al., 2004), where a threshold is optimized on the corresponding AER of the baseline model (i.e., HMM+*sym* or IBM Model 4+*sym*)[7].

Table 2 shows the results of word alignment evaluations[8], where *none* denotes that the model has no constraint. In KFTT and BTEC Corpus, *itg* achieved significant improvement against *sym* and *none* on IBM Model 4 ($p \leq 0.05$)[9]. However, in the Hansard Corpus, *itg* shows no improvement against *sym*. This indicates that capturing structural coherence by *itg* yields a significant benefit to word alignment in a linguistically different language pair such as Ja-En. For example, some function words appear more than once in both a source and target sentence, and they are not symmetrically aligned with each other, especially in regards to the Ja-En language pair. Although the baseline methods tend to be unable to align such long-distance word pairs, the proposed method can correctly catch them because *itg* can determine the relation of long-distance words. We discuss more details about the effectiveness of the ITG constraint in Section 4.1.

### 3.2 Translation Evaluation

We measured translation performance with BLEU (Papineni et al., 2002). All language models are 5-gram and trained using SRILM (Stolcke and others, 2002) on target side sentences in the training data. When extracting phrases, we apply the method proposed by Matusov et al. (2004), where many-to-many alignments are generated based on the averages of the posterior probabilities from two directional models[10].

We used the Moses phrase-based SMT systems (Koehn et al., 2007) for decoding. We set the distortion-limit parameter to infinite[11], and other pa-

[5] BTEC Corpus is a subset of IWSLT 2007. To uniform tokenization, we retokenized all Japanese sentences both in IWSLT 2007 and BTEC Corpus using ChaSen (Asahara and Matsumoto, 2000).

[6] Since there exists no distinction for sure-possible alignments in the KFTT and BTEC data sets, we treat all alignments of them as sure alignments.

[7] We tried values from 0.1 to 1.0 at an interval of 0.1.

[8] The values in bold indicate the best score.

[9] The statistical significance test was performed by the paired bootstrap resampling (Koehn, 2004).

[10] The posterior thresholds were decided in the same way as the word alignment evaluation.

[11] This setting is generally used for Ja-En translation tasks (Murakami et al., 2007).

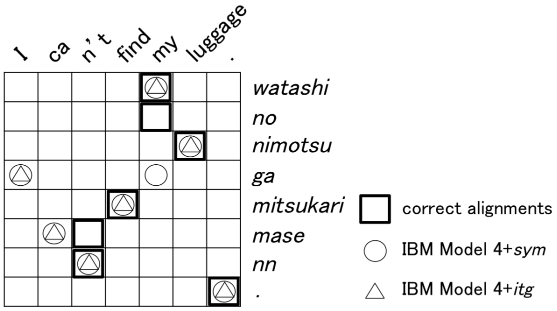

**Figure 1:** Word alignment examples on the BTEC corpus.

rameters as default settings. Parameter tuning was conducted by 100-best batch MIRA (Cherry and Foster, 2012) with 25 iterations.

Table 3 shows the average BLEU of five different tunings[12]. In both KFTT and IWSLT 2007, *itg* achieved significant improvement against both *none* and *sym* on HMM model. On IBM Model4, *itg* significantly outperforms *none* and is comparable to *sym* in KFTT, while *itg* significantly outperforms *sym* and is comparable to *none* in IWSLT 2007.

# 4 Discussion

## 4.1 Effects of ITG Constraints on Word Alignment and Translation

We discuss the effect of our ITG constraint on word alignment and machine translation. As described in Section 2, the ITG constraint is imposed in the E-step of the EM algorithm, not in decoding steps. Therefore, for the sentences that are not contained in the training corpus, the word alignments are calculated using the emission, transition and fertility tables trained with the constraint. It means that the effects of the constraint are implicitly reflected in the alignment results. On the other hand, the effects of the constraint are directly reflected in the machine translation results because the phrase tables are extracted from the posterior probabilities calculated in training steps. Therefore, our ITG constraint has a potential to achieve a large improvement of machine translation performance relative to an improvement of alignment performance, such as IBM Model 4+*itg* vs. IBM Model 4+*sym* on the BTEC corpus. We would like to improve our model by imposing our ITG constraint on decoding steps in future.

## 4.2 Comparison between Symmetric and ITG Constraint

In KFTT, *itg* is comparable to *sym* on IBM Model 4 in machine translation; however, *itg* achieved significant improvement in terms of word alignment, which follows the previous reports that better word alignment does not always result in better translation (Ganchev et al., 2008; Yang et al., 2013). On the other hand, in BTEC, *itg* outperforms *sym* both on word alignment and machine translation. Figure 1 shows that IBM Model 4+*sym* often generates wrong gappy alignments such as "*ga* (Ja)-I (En)" and "*ga* (Ja)-my (En)". These wrong alignments disturb the phrase extraction, because excessively long phrase pairs are extracted by bridging the gaps in wrong alignments or simply no phrase pairs are extracted from wrong gappy alignments. Consequently, the phrase table generated by IBM Model 4+*sym* tend to be sparse and contain longer phrase pairs than the one generated by IBM Model 4+*itg*.

# 5 Conclusions

We have proposed a novel alignment method that uses an ITG constraint based on bracketing ITG parse trees as a soft constraint of the posterior regularization framework. Due to the ITG constraint, the proposed method can symmetrize two directional alignments based on their structural coherence. Our evaluations have shown that the proposed ITG constraint significantly improves the baseline word alignment performance on the Ja-En KFTT and BTEC corpus, and significantly improves, or at least keeps, the baseline machine translation performance of KFTT and the Ja-En IWSLT 2007 task. This indicates that the proposed method yields a significant benefit to linguistically different language pairs.

In future work, we plan to incorporate a phrasal ITG (Cherry and Lin, 2007) instead of a bracketing ITG to efficiently handle many-to-many alignments.

[12] The values in bold represent the best score, and † indicates that the comparisons are not significant over the corresponding model (i.e., HMM+*itg* or IBM Model 4+*itg*) according to the bootstrap resampling test ($p \leq 0.05$). We used multeval (Clark et al., 2011) for significance testing.

## References

Masayuki Asahara and Yuji Matsumoto. 2000. Extended models and tools for high-performance part-of-speech tagger. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 21–27. Association for Computational Linguistics.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada, June. Association for Computational Linguistics.

Colin Cherry and Dekang Lin. 2007. Inversion Transduction Grammar for Joint Phrasal Translation Modeling. In *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 17–24, Rochester, New York, April. Association for Computational Linguistics.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Goh Chooi-Ling, Watanabe Taro, Yamamoto Hirofumi, and Sumita Eiichiro. 2010. Constraining a generative word alignment model with discriminative output.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA, June. Association for Computational Linguistics.

John DeNero and Dan Klein. 2007. Tailoring Word Alignments to Syntactic Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 17–24, Prague, Czech Republic, June. Association for Computational Linguistics.

Cameron S Fordyce. 2007. Overview of the IWSLT 2007 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation 2007*, pages 1–12.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Kuzman Ganchev, João V. Graça, and Ben Taskar. 2008. Better Alignments = Better Translations? In *Proceedings of ACL-08: HLT*, pages 986–993, Columbus, Ohio, June. Association for Computational Linguistics.

Kuzman Ganchev, Joao Graca, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049.

Joao V Graça, Kuzman Ganchev, and Ben Taskar. 2008. Expectation Maximization and Posterior Constraints. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 569–576. Curran Associates, Inc.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

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 the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Shuhei Kondo, Kevin Duh, and Yuji Matsumoto. 2013. Hidden Markov Tree Model for Word Alignment. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 503–511, Sofia, Bulgaria, August. Association for Computational Linguistics.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by Agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York City, USA, June. Association for Computational Linguistics.

Xiaodong Liu, Kevin Duh, and Yuji Matsumoto. 2013. Topic models + word alignment = a flexible framework for extracting bilingual dictionary from comparable corpus. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 212–221, Sofia, Bulgaria, August. Association for Computational Linguistics.

Evgeny Matusov, Richard Zens, and Hermann Ney. 2004. Symmetric Word Alignments for Statistical Machine Translation. In *Proceedings of COLING 2004, the 20th International Conference on Compu-*

*tational Linguistics*, pages 219–225, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Rada Mihalcea and Ted Pedersen. 2003. An Evaluation Exercise for Word Alignment. In Rada Mihalcea and Ted Pedersen, editors, *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10.

Jin'ichi Murakami, Tokuhisa Masato, and Satoru Ikehara. 2007. Statistical machine translation using large j/e parallel corpus and long phrase tables. In *Proceedings of the International Workshop on Spoken Language Translation 2007*, pages 151–155.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Markus Saers, Karteek Addanki, and Dekai Wu. 2012. From Finite-State to Inversion Transductions: Toward Unsupervised Bilingual Grammar Induction. In *Proceedings of COLING 2012, the 24th International Conference on Computational Linguistics*, pages 2325–2340, Mumbai, India, December. The COLING 2012 Organizing Committee.

Andreas Stolcke et al. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286, November.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 147–152.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational Linguistics-Volume 2*, pages 836–841. Association for Computational Linguistics.

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

Xinyan Xiao, Deyi Xiong, Yang Liu, Qun Liu, and Shouxun Lin. 2012. Unsupervised Discriminative Induction of Synchronous Grammar for Machine Translation. In *Proceedings of COLING 2012, the 24th International Conference on Computational Linguistics*, pages 2883–2898, Mumbai, India, December. The COLING 2012 Organizing Committee.

Nan Yang, Shujie Liu, Mu Li, Ming Zhou, and Nenghai Yu. 2013. Word Alignment Modeling with Context Dependent Deep Neural Network. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 166–175, Sofia, Bulgaria, August. Association for Computational Linguistics.

Hao Zhang and Daniel Gildea. 2004. Syntax-Based Alignment: Supervised or Unsupervised? In *Proceedings of COLING 2004, the 20th International Conference on Computational Linguistics*, pages 418–424, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Hao Zhang and Daniel Gildea. 2005. Stochastic Lexicalized Inversion Transduction Grammar for Alignment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 475–482, Ann Arbor, Michigan, June. Association for Computational Linguistics.