Unsupervised Word Alignment Using Frequency Constraint in Posterior Regularized EM

Hidetaka Kamigaito[†], Taro Watanabe^{††}, Hiroya Takamura^{†††},
Manabu Okumura^{†††} and Eiichiro Sumita^{†††}

Generative word alignment models, such as IBM Models, are restricted to one-to-many alignment, and cannot explicitly represent many-to-many relationships in bilingual texts. The problem is partially solved either by introducing heuristics or by agreement constraints such that two directional word alignments agree with each other. However, this constraint cannot take into account the grammatical difference of language pairs. In particular, function words are not trivial to align for grammatically different language pairs, such as Japanese and English. In this paper, we focus on the posterior regularization framework (Ganchev, Graca, Gillenwater, and Taskar 2010) that can force two directional word alignment models to agree with each other during training, and propose new constraints that can take into account the difference between function words and content words. We discriminate a function word and a content word using word frequency in the same way as done by Setiawan, Kan, and Li (2007). Experimental results show that our proposed constraints achieved better alignment qualities on the French-English Hansard task and the Japanese-English Kyoto free translation task (KFTT) measured by AER and F-measure. In translation evaluations, we achieved statistically significant gains in BLEU scores in the Japanese-English NTCIR10 task and Spanish-English WMT06 task.

Key Words: *Statistical Machine Translation, Unsupervised Word Alignment, Posterior Regularization Framework, Constrained EM*

1 Introduction

Word alignment is an important component in statistical machine translation (SMT). For instance phrase-based SMT (Koehn, Och, and Marcu 2003) is based on the concept of phrase pairs from bilingual data with word alignment annotation. Similarly, the model for hierarchical phrase-based SMT is built from exhaustively extracted phrases that are, in turn, heavily reliant on word alignment.

[†] Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

^{††} Google Inc. This work was mostly done while the second author was affiliated with National Institute of Information and Communications Technology

^{†††} Tokyo Institute of Technology, Precision and Intelligence Laboratory

^{††††} National Institute of Information and Communication Technology

The Generative word alignment models, such as the IBM Models (Brown, Pietra, Pietra, and Mercer 1993) and HMM (Vogel, Ney, and Tillmann 1996), are popular methods for automatically aligning bilingual texts. However, they are restricted to represent one-to-many correspondence of each word. To resolve this weakness, various symmetrization methods are proposed. Och and Ney (2003) and Koehn et al. (2003) propose various heuristic methods to combine two directional models to represent many-to-many relationships. As an alternative to heuristic methods, filtering methods employ a threshold to control the trade-off between precision and recall based on a score estimated from the posterior probabilities from two directional models. Matusov, Zens, and Ney (2004) proposed arithmetic means of two models as a score for the filtering, whereas Liang, Taskar, and Klein (2006) reported better results using geometric means. The joint training method (Liang et al. 2006) enforces agreement between two directional models. Posterior regularization (Ganchev et al. 2010) is an alternative agreement method which directly encodes agreement during training. DeNero and Macherey (2011) and Chang, Rush, DeNero, and Collins (2014) also enforce agreement during inference, not training. However, these agreement models do not take into account the difference in language pairs, which is crucial for linguistically different language pairs, such as Japanese and English: although content words may be aligned with each other by introducing some agreement constraints, function words are difficult to align. This is because the function words have many roles in a single word. Nakazawa and Kurohashi (2012) reported larger gains against a heuristic baseline by a special treatment for function words. However their method relied on dependency analysis for discriminating function words.

We focus on the posterior regularization framework and improve upon the previous work by proposing new constraint functions that take into account the difference in languages in terms of content words and function words. In particular, we concentrate on a method which does not rely on deeper syntactic analysis and differentiate content words and function words by the frequency in bilingual data, following Setiawan et al. (2007). In this work, content words and function words are discriminated based on the frequency on a given data by treating high frequent words as function words and low frequent words as content words. The difference of a function word and a content word is directly reflected by our proposed three constraints: $f2f$ constraint rewards posterior probability of alignments between a function word and a function word, $c2c$ constraint rewards posterior probability of alignments between two content words, $f2c$ constraint penalize posterior probability of alignments between a function word and a content word. These three fine grained agreement constraints are used as a symmetric constraint for the posterior regularization framework (Ganchev et al. 2010).

Experimental results show that the proposed methods achieved better alignment qualities on

the French-English Hansard data and the Japanese-English Kyoto free translation task (KFTT) measured by AER and F-measure. In translation evaluations, we achieved statistically significant gains in BLEU scores in the NTCIR10 Japanese-English task and WMT06 Spanish-English task.

The remainder of this paper is organized as follows. Section 2 describes the related work of one-directional and bidirectional word alignment. Section 3 describes the summary of Posterior Regularization Framework and Ganchev’s symmetric constraint. Section 4 describes our proposed two constraint functions, matching and mismatching constraint functions. Section 5 describes the comparison between previous symmetric function and our proposed constraint functions against various language pairs. Section 6 describes the analysis and discussion of the results of Experiments. Section 7 describes the conclusion and future work. This paper is an extension of our prior work with more details and experiments (Kamigaito, Watanabe, Takamura, and Okumura 2014).

2 Related Work

IBM Models (Brown et al. 1993) and HMM (Vogel et al. 1996) are popular generative models for representing word alignment in bilingual texts. One of the shortcomings of these models is the one-to-many restriction, in which each target word can be aligned to just a single source word. As a result, Och and Ney (2003) and Koehn et al. (2003) presented various symmetrization methods in which Viterbi alignments from two directional models are heuristically combined to yield many-to-many alignments. Their heuristic methods work mostly fine though they cannot clearly take into account the posterior probability of word alignments on various granularities.

Some generative models can directly induce phrasal structures without one-to-many word alignment constraints (DeNero, Bouchard-Côté, and Klein 2008; Neubig, Watanabe, Sumita, Mori, and Kawahara 2011b). These methods can learn compact phrase pairs with comparative translation quality against heuristic extraction methods. However these methods take long time for inference. From these reasons, heuristic extractions are still used in various NLP tasks.

As an alternative to heuristic methods, filtering methods employ a threshold to control the trade-off between precision and recall based on a score estimated from the posterior probabilities from two directional models. Matusov et al. (2004) proposed arithmetic means of two models as a score for the filtering, whereas Liang et al. (2006) reported better results using geometric means. Both methods can generate bidirectional alignments based on posterior probability with various granularities. However, these models still separately learning the two directional alignment models. For this reason, both models sometimes misalign the bidirectional alignments.

Instead of training two models independently, Liang et al. (2006) propose simultaneous training of two models, in a parallel manner, by maximizing the product of the two models during the M-step in the EM-algorithm. The E-step is intractable even with a simpler variant of IBM Models, since summing over all possible alignment space involves enumeration of all bidirectional alignment sets. To solve the problem, they use a heuristic approximation by simply considering the product of two models. Model-Based Aligner Combination (DeNero and Macherey 2011) enforces agreement during the inference step. Since the inference over two directional models is intractable, they employ dual decomposition as an approximated inference. Constrained Viterbi Relaxation (Chang et al. 2014) allows even faster inference with guaranteed optimality. Ganchev et al. (2010) introduce a posterior regularization framework that involves regularization of posterior probabilities to match a certain constraint during the E-step in the EM-algorithm. Using the approximation through dual decomposition, this framework can use various constraints as a function to incorporate prior knowledge. They propose agreement constraints, or features, of two word alignment models, which can correct the posterior probability of a word alignment during training.

Ganchev et al. (2010) and DeNero and Macherey (2011) are similar in the spirit of generating bidirectionally agreed word alignment. However, they differ in that Ganchev et al. (2010) enforces agreement of two directional models using dual decomposition in the training step, whereas that of DeNero and Macherey (2011) is in the inference step.

Nakazawa and Kurohashi (2012) proposed a method to improve word alignment accuracies by differentiating function words using dependency trees. However, their method strongly relies on the accuracies of dependency trees and such trees may not be available to under-resourced languages.

3 Statistical Word Alignment with Posterior Regularization Framework

3.1 Overview

Posterior regularization framework is an instance of constrained EM algorithm in which prior knowledge can be easily introduced as a soft constraint. In E-step of EM algorithm, posterior regularization framework modify the posterior probability such that it may not violate a given constraint. In word alignment task, a symmetric constraint is used to enforce two independent directional alignment models to agree with each other by moving the posterior probabilities for every alignment between a source word and a target word. Figure 1 shows an example of

Fig. 1 Posterior probability of bidirectional agreed alignments

symmetric constraint under posterior regularization framework for a Japanese-English¹ sentence pair, “*kuroi inu ga ni hiki ita.*”² and “There were two black dogs.” with each bar representing the posterior probability for each word alignment. The mass in the top left are the posterior probabilities for English-to-Japanese alignment model denoted as $\vec{p}$, and the mass in the top right are those for Japanese-to-English alignment model denoted as $\overleftarrow{p}$. The two distributions are transferred as new posterior probabilities after agreement under posterior regularization denoted by q . The probability of alignment between Japanese word “*hiki*” and English word “two” is lower than that of Japanese word “*hiki*” and English word “dogs” in Japanese-to-English alignment. On the other hand, the probability of alignment between Japanese word “*hiki*” and English word “two” is lower than that of Japanese word “*hiki*” and English word “There” in English-to-Japanese alignment. The symmetric constraint enforces weighted arithmetic means of the probability of both Japanese-to-English and English-to-Japanese alignments as a posterior probability of bidirectional alignment during training. Through the training with symmetric constraint, the probability of alignment between Japanese word “*hiki*” and English word “two” is higher than that of “*hiki*” and “dogs” and “*hiki*” and “There” in the agreed bidirectional alignment.

¹ This notation denotes that the source language is Japanese and the target language is English. The notation of Japanese-to-English is the same.

² Japanese words have kanji, hiragana, katakana and alphabet styles. For simple expression, in this paper, all Japanese words are written in italic roman style of its pronunciation.

3.2 Symmetric Constraint

In this section, we explain the detail of symmetric constraint under posterior regularization framework. Given a bilingual sentence $\mathbf{x} = (\mathbf{x}^s, \mathbf{x}^t)$ where $\mathbf{x}^s$ and $\mathbf{x}^t$ denote a source and target sentence, respectively, and the bilingual sentence is aligned by a word alignment denoted by $\mathbf{y}$. In particular, $y_{i,j} = 1$, if a source word x_i^s is aligned to a target word x_j^t (otherwise $y_{i,j} = 0$) where we define the index of $\mathbf{x}^t$, $\mathbf{x}^s$ as i, j ($1 \leq i \leq |\mathbf{x}^t|, 1 \leq j \leq |\mathbf{x}^s|$). We represent posterior probabilities from two directional word alignment models as $\vec{p}_\theta(\mathbf{y}|\mathbf{x})$ and $\overleftarrow{p}_\theta(\mathbf{y}|\mathbf{x})$ with each arrow indicating a particular direction, and use θ to denote the parameters of the models. To explain the directional models, we additionally denote a directional word alignment as $\mathbf{a}$. Note that an alignment $a_i = j$ is equivalent to $y_{i,j} = 1$. For instance, $\mathbf{a}$ is a set of the alignment from $\mathbf{x}^s$ to $\mathbf{x}^t$ under the model of $p(\mathbf{x}^t, \mathbf{y}|\mathbf{x}^s)$ in which each target word is aligned with at most one source word. In the case of IBM Model 1, let a_i is an alignment of target word x_i^t and p_t is a word translation probability, the model is represented as follows:

$$p(\mathbf{x}^t, \mathbf{y}|\mathbf{x}^s) = \prod_i \frac{1}{|\mathbf{x}^s| + 1} p_t(x_i^t | x_{a_i}^s). \quad (1)$$

Note that Equation (1) assumes a special NULL symbol located at the 0th position in $\mathbf{x}^s$ denoted as x_0^s . The model for the inverse direction $p(\mathbf{x}^s, \mathbf{y}|\mathbf{x}^t)$ is defined similarly using a special NULL symbol for x_0^t . The posterior probabilities are defined as follows:

$$\vec{p}(\mathbf{y}|\mathbf{x}) = \prod_i \frac{p_t(x_i^t | x_{a_i}^s)}{\sum_{j'} p_t(x_i^t | x_{j'}^s)}, \quad (2)$$

where j' ($0 \leq j' \leq |\mathbf{x}^s|$) is an index of $\mathbf{x}^s$. We denote a set of alignments generated by a source-to-target model as $\vec{\mathbf{Y}}$ and by a target-to-source alignment model as $\overleftarrow{\mathbf{Y}}$. Herein, we assume that the posterior probability for both directional alignment is zero (i.e., $\vec{p}(\mathbf{y}|\mathbf{x}) = 0$ for $\mathbf{y} \notin \vec{\mathbf{Y}}$), in which an alignment may be possible in one direction, but not in the other due to the one-to-many restriction. Given the two directional models, we define a symmetric feature for each target/source position pair, i, j as follows:

$$\phi_{i,j}(\mathbf{x}, \mathbf{y}) = \begin{cases} +1 & \overleftarrow{Z}(i, j), \\ -1 & \vec{Z}(i, j), \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

where $\overleftarrow{Z}(i, j) = (\mathbf{y} \in \overleftarrow{\mathbf{Y}}) \wedge (y_{i,j} = 1)$ and $\vec{Z}(i, j) = (\mathbf{y} \in \vec{\mathbf{Y}}) \wedge (y_{i,j} = 1)$. The feature assigns 1 to an alignment of $\overleftarrow{\mathbf{Y}}$, but assigns -1 to an element of $\vec{\mathbf{Y}}$. As a result, if a word pair i, j is aligned with equal posterior probabilities in two directions, the expectation of the feature value

will be zero. We define a joint model that combines two directional models by arithmetic means:

$$p_{\theta}(\mathbf{y}|\mathbf{x}) = \frac{1}{2} \vec{p}_{\theta}(\mathbf{y}|\mathbf{x}) + \frac{1}{2} \overleftarrow{p}_{\theta}(\mathbf{y}|\mathbf{x}). \quad (4)$$

Under the posterior regularization framework, we instead use q parameterized by λ for each bilingual data $\mathbf{x}$ as follows (Ganchev et al. 2010):

$$\begin{aligned} q_{\lambda}(\mathbf{y}|\mathbf{x}) &= \frac{\vec{p}_{\theta}(\mathbf{y}|\mathbf{x}) + \overleftarrow{p}_{\theta}(\mathbf{y}|\mathbf{x})}{2} \cdot \frac{\exp\{-\lambda \cdot \phi(\mathbf{x}, \mathbf{y})\}}{Z} \\ &= \frac{\vec{q}(\mathbf{y}|\mathbf{x})^{\frac{Z_{\vec{q}}}{\vec{p}_{\theta}(\mathbf{x})}} + \overleftarrow{q}(\mathbf{y}|\mathbf{x})^{\frac{Z_{\overleftarrow{q}}}{\overleftarrow{p}_{\theta}(\mathbf{x})}}}{2Z}, \\ Z &= \frac{1}{2} \left(\frac{Z_{\vec{q}}}{\vec{p}_{\theta}(\mathbf{x})} + \frac{Z_{\overleftarrow{q}}}{\overleftarrow{p}_{\theta}(\mathbf{x})} \right), \\ \vec{q}(\mathbf{y}|\mathbf{x}) &= \frac{1}{Z_{\vec{q}}} \vec{p}_{\theta}(\mathbf{y}, \mathbf{x}) \exp\{-\lambda \cdot \phi(\mathbf{x}, \mathbf{y})\}, \\ Z_{\vec{q}} &= \sum_{\mathbf{y}} \vec{p}_{\theta}(\mathbf{y}, \mathbf{x}) \exp\{-\lambda \cdot \phi(\mathbf{x}, \mathbf{y})\}, \\ \overleftarrow{q}(\mathbf{y}|\mathbf{x}) &= \frac{1}{Z_{\overleftarrow{q}}} \overleftarrow{p}_{\theta}(\mathbf{y}, \mathbf{x}) \exp\{-\lambda \cdot \phi(\mathbf{x}, \mathbf{y})\}, \\ Z_{\overleftarrow{q}} &= \sum_{\mathbf{y}} \overleftarrow{p}_{\theta}(\mathbf{y}, \mathbf{x}) \exp\{-\lambda \cdot \phi(\mathbf{x}, \mathbf{y})\}, \end{aligned} \quad (5)$$

to satisfy the following constraint:

$$\forall i, \forall j, \vec{q}_{\lambda_{i,j}}(y_{i,j}=1|\mathbf{x}) - \overleftarrow{q}_{\lambda_{i,j}}(y_{i,j}=1|\mathbf{x}) = 0. \quad (6)$$

In the E-step of EM-algorithm, we employ q_{λ} instead of p_{θ} to accumulate fractional counts for its use in the M-step. λ in the q_{λ} controls the two directional posterior probabilities $\vec{p}_{\theta}$ and $\overleftarrow{p}_{\theta}$ for a symmetrical word alignment. For example, a large $\lambda_{i,j}$ increases $\vec{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$ and decreases $\overleftarrow{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$. Conversely, a small $\lambda_{i,j}$ decreases $\vec{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$ and increases $\overleftarrow{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$. When the constraint (6) is satisfied, $\vec{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$ and $\overleftarrow{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$ are symmetrical. λ is efficiently estimated by the gradient ascent for each bilingual sentence $\mathbf{x}$. Note that posterior regularization is performed during parameter estimation, and not during testing.

4 Posterior Regularization with Frequency Constraint

The symmetric constraint method represented in Equation (3) assumes a strong one-to-one relation for any word, and does not take into account the divergence in language pairs. For linguistically different language pairs, such as Japanese-English, content words may be easily

Fig. 2 Example of a function-to-content wrong alignment

aligned one-to-one, but function words are not always aligned together. Figure 2 is an example of a wrong alignment between an English function word “There” to a Japanese content word “*hiki*”. Note that Japanese word “*hiki*” is a counter word. These errors may impact the end-to-end translation qualities.

In order to solve the problem, we improve Ganchev’s symmetric constraint so that it can consider the difference between content words and function words in each language. In particular, we follow the frequency-based idea of Setiawan et al. (2007) that discriminates content words and function words by their frequencies. We propose constraint features that take into account the difference between content words and function words, determined by a frequency threshold.

4.1 Discriminating Content Words and Function Words

Setiawan et al. (2007) discriminated content words and function words by their frequency using a threshold, and successfully employed the knowledge for a better hierarchical grammar. Intuitively, the frequency for function words is high whereas that for content words is low. However, they employed a fixed constant as the threshold parameter, which is largely influenced by the data size. Bisazza and Federico (2012) solved this problem by introducing a threshold determined by the ratio of very frequent words. In particular, the threshold th is determined as the maximum frequency that satisfies the following condition:

$$\frac{\sum_{w \in (freq(w) > th)} freq(w)}{\sum_{w \in all} freq(w)} > r. \quad (7)$$

Here, we empirically set $r = 0.5$. If a word has higher frequency than the threshold th , we treat the word as a function word, otherwise, we treat the word as a content word.

4.2 Mismatching Constraint

First, we propose a mismatching constraint that penalizes word alignment between content words and function words by decreasing the corresponding posterior probabilities. This constraint function penalizes alignment for *function to content* or *content to function* word matching, namely *f2c*. Figure 3 is an example of posterior probability of an alignment between an English function word “There” to a Japanese content word “hiki”. In this example, the posterior probability of Japanese-to-English direction is lower than the other, and thus the posterior probability is lowered for the bidirectional model under the mismatching constraint. Using functions $\overleftarrow{Z}(i, j) = (\mathbf{y} \in \overleftarrow{\mathbf{Y}}) \wedge (\overleftarrow{y}_{i,j} = 1)$ and $\overrightarrow{Z}(i, j) = (\mathbf{y} \in \overrightarrow{\mathbf{Y}}) \wedge (\overrightarrow{y}_{i,j} = 1)$ both defined in Equation (3), and $F2C(x_i^t, x_j^s) = ((x_i^t \in \mathcal{F}^t \wedge x_j^s \in \mathcal{C}^s) \vee (x_i^t \in \mathcal{C}^t \wedge x_j^s \in \mathcal{F}^s))$, the *f2c* constraint function for a word pair (x_i^t, x_j^s) is formally defined as follows:

$$\phi_{i,j}^{f2c}(\mathbf{x}, \mathbf{y}) = \begin{cases} +1 & \overleftarrow{Z}(i, j) \wedge F2C(x_i^t, x_j^s) \wedge (\delta_{i,j}(\mathbf{x}, \mathbf{y}) < 0), \\ 0 & \overleftarrow{Z}(i, j) \wedge F2C(x_i^t, x_j^s) \wedge (\delta_{i,j}(\mathbf{x}, \mathbf{y}) > 0), \\ 0 & \overrightarrow{Z}(i, j) \wedge F2C(x_i^t, x_j^s) \wedge (\delta_{i,j}(\mathbf{x}, \mathbf{y}) < 0), \\ -1 & \overrightarrow{Z}(i, j) \wedge F2C(x_i^t, x_j^s) \wedge (\delta_{i,j}(\mathbf{x}, \mathbf{y}) > 0), \\ +1 & \overrightarrow{Z}(i, j) \wedge \neg(F2C(x_i^t, x_j^s)) \wedge (\delta_{i,j}(\mathbf{x}, \mathbf{y}) < 0), \\ -1 & \overleftarrow{Z}(i, j) \wedge \neg(F2C(x_i^t, x_j^s)) \wedge (\delta_{i,j}(\mathbf{x}, \mathbf{y}) > 0), \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

where $\delta_{i,j}(\mathbf{x}, \mathbf{y}) = \overrightarrow{p}_{\theta}(y_{i,j} = 1|\mathbf{x}) - \overleftarrow{p}_{\theta}(y_{i,j} = 1|\mathbf{x})$ is the difference in the posterior probabilities between the source-to-target and the target-to-source alignments. $\mathcal{C}^s$ and $\mathcal{C}^t$ represent content words in the source sentence and target sentence, respectively. Similarly, $\mathcal{F}^s$ and $\mathcal{F}^t$ are function words in the source and target sentence, respectively. Intuitively, when there exists a mismatch in content word and function word for a word pair (x_i^t, x_j^s) , the constraint function returns a non-zero

Fig. 3 Posterior probability of bidirectional alignment generated by mismatching constraint

value for the model with the highest posterior probability. When coupled with the constraint such that the expectation of the feature value is zero, the constraint function decreases the posterior probability of the highest direction and discourages agreement with each other for matching with a content word and a function word. Similarly the symmetric constraint, λ controls the two directional posterior probabilities $\vec{q}_\lambda$ and $\overleftarrow{q}_\lambda$ for penalizing word alignment between content words and function words. For example, when (i, j) is a word pair between a content word and a function word, and $\overleftarrow{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$ is larger than $\vec{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$, the first and the third rules would fire such that increased $\lambda_{i,j}$ might penalize $\overleftarrow{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$ until it reaches $\vec{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$. Similarly, the second and fourth rules would fire when $\vec{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$ is larger than $\overleftarrow{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$, which might lead to decreased $\lambda_{i,j}$ to penalize $\vec{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$. If $y_{i,j}$ does not satisfy $F2C(x_i^t, x_j^s)$, our constraint $\phi_{i,j}^{f2c}$ operates similarly to $\phi_{i,j}^{sym}$ according to the last three rules.

4.3 Matching Constraint

In contrast to the mismatching constraint, our second constraint function rewards alignment for *function to function* word matching, namely *f2f* and for *content to content* word matching, namely *c2c*. Figure 4 is an example of posterior probability of an alignment between an English content word “dogs” to a Japanese content word “inu”. In this example, the posterior probability of English-to-Japanese direction is higher than the other; we prefer the English-to-Japanese direction as the posterior probability for the bidirectionally agreed model. Using functions $\overleftarrow{Z}(i, j) = (\mathbf{y} \in \overleftarrow{\mathbf{Y}}) \wedge (\overleftarrow{y}_{i,j} = 1)$ and $\overrightarrow{Z}(i, j) = (\mathbf{y} \in \overrightarrow{\mathbf{Y}}) \wedge (\overrightarrow{y}_{i,j} = 1)$ both defined in Equation (3), and $C2C(x_i^t, x_j^s) = (x_i^t \in \mathcal{C}^t \wedge x_j^s \in \mathcal{C}^s)$, the *c2c* constraint function for a word pair (x_i^t, x_j^s) is formally defined as follows:

Fig. 4 Posterior probability of bidirectional alignment generated by matching constraint

$$\phi_{i,j}^{c2c}(\mathbf{x}, \mathbf{y}) = \begin{cases} 0 & \overleftarrow{Z}(i, j) \wedge C2C(x_i^t, x_j^s) \wedge (\delta_{i,j}(\mathbf{x}, \mathbf{y}) < 0), \\ +1 & \overleftarrow{Z}(i, j) \wedge C2C(x_i^t, x_j^s) \wedge (\delta_{i,j}(\mathbf{x}, \mathbf{y}) > 0), \\ -1 & \overrightarrow{Z}(i, j) \wedge C2C(x_i^t, x_j^s) \wedge (\delta_{i,j}(\mathbf{x}, \mathbf{y}) < 0), \\ 0 & \overrightarrow{Z}(i, j) \wedge C2C(x_i^t, x_j^s) \wedge (\delta_{i,j}(\mathbf{x}, \mathbf{y}) > 0), \\ +1 & \overleftarrow{Z}(i, j) \wedge \neg(C2C(x_i^t, x_j^s)) \wedge (\delta_{i,j}(\mathbf{x}, \mathbf{y}) < 0), \\ -1 & \overrightarrow{Z}(i, j) \wedge \neg(C2C(x_i^t, x_j^s)) \wedge (\delta_{i,j}(\mathbf{x}, \mathbf{y}) > 0), \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

This constraint function returns a non-zero value for a word pair (x_i^t, x_j^s) when they are content words. As a result, pairs of content words are encouraged to agree with each other, but the other pairs. The *function to function* word matching function $f2f$ can be defined similarly by replacing C^s and C^t by $\mathcal{F}^s$ and $\mathcal{F}^t$, respectively. Likewise, the *function to content* word matching function $f2c$ is defined by considering the matching of content words and function words in two languages. Similarly the symmetric constraint, λ controls the two directional posterior probabilities $\overrightarrow{q}_\lambda$ and $\overleftarrow{q}_\lambda$ for encouraging word alignment between content words and function words. For example, when (i, j) is a word pair between a content word and a content word, and $\overleftarrow{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$ is larger than $\overrightarrow{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$, the first and the third rules would fire such that decreased $\lambda_{i,j}$ might increase $\overrightarrow{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$ until it reaches $\overleftarrow{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$. Similarly, the second and fourth rules would fire when $\overleftarrow{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$ is larger than $\overrightarrow{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$, which might lead to increased $\lambda_{i,j}$ to increase $\overleftarrow{q}_{\lambda_{i,j}}(y_{i,j} = 1|\mathbf{x})$. If $y_{i,j}$ does not satisfy $C2C(x_i^t, x_j^s)$, our constraint $\phi_{i,j}^{c2c}$ operates similarly to $\phi_{i,j}^{sym}$ according to the last three rules.

5 Experiment

5.1 Experimental Setup

We used French-English Hansard Corpus and Japanese-English Kyoto free translation task (KFTT) for word alignment evaluation. In the machine translation evaluation, we used the German-English, Spanish-English and French-English WMT06 with europarl-v7 training data set and Japanese-English NTCIR10. The KFTT (Neubig 2011) is derived from Japanese Wikipedia articles related to Kyoto, which is translated into English by expert translators. NTCIR10 comes from patent data employed in a machine translation shared task (Goto, Chow, Lu, Sumita, and Tsou 2013). The statistic of these data is presented in Table 1 and 2. Sentences of over 40 words on both source and target sides are removed for training alignment models. We use cicada³ for

³ <https://github.com/tarowatanabe/cicada>

Table 1 The statistics of the data sets used in word alignment evaluation

Corpus		train			test		
		sentence	word	vocab	sentence	word	vocab
Hansard fr-en	fr	1.13 M	23.3 M	78.1 k	447	7.76 k	1.92 k
	en		19.8 M	57.3 k		7.02 k	1.69 k
KFTT ja-en	ja	329.88 k	6.08 M	26.8 k	582	14.4 k	2.57 k
	en		5.91 M	24.3 k		12.6 k	2.65 k

Table 2 The statistics of the data sets used in translation evaluation

Corpus		train			dev			test		
		sentence	word	vocab	sentence	word	vocab	sentence	word	vocab
europarl-v7 de-en	de	1.53 M	31.32 M	289.78 k	2.00 k	55.12 k	8.79 k	2.00 k	55.53 k	8.81 k
	en		32.83 M	90.81 k		58.76 k	6.12 k		59.42 k	6.03 k
europarl-v7 es-en	es	1.52 M	32.96 M	133.78 k	2.00 k	60.47 k	7.68 k	2.00 k	61.72 k	7.72 k
	en		31.85 M	89.15 k		58.76 k	6.12 k		59.42 k	6.03 k
europarl-v7 fr-en	fr	1.50 M	33.71 M	107.67 k	2.00 k	64.33 k	7.29 k	2.00 k	65.10 k	7.24 k
	en		30.72 M	87.61 k		58.76 k	6.12 k		59.42 k	6.03 k
NTCIR10 ja-en	ja	2.02 M	53.40 M	113.86 k	2.00 k	73.01 k	4.38 k	8.60 k	333.82 k	10.40 k
	en		49.37 M	182.55 k		67.35 k	5.04 k		310.18 k	12.70 k

training the HMM and IBM Model 4 with our proposed methods. Training is bootstrapped from IBM Model 1, followed by HMM and IBM Model 4. In both word alignment and translation evaluation, we used filtering method (Liang et al. 2006) with threshold parameter 0.1 for detailed analysis. This threshold value were selected from 0.1 to 0.9 with interval 0.1 through preliminary tests.

5.2 Word Alignment Evaluation

We measure the impact of our proposed methods on the quality of word alignment measured by AER (Och and Ney 2003) and F-measure. We used only sure alignment for calculating precision, recall and F-measure (Fraser and Marcu 2007). For the detailed analysis, we also evaluated word alignments by focusing on particular instances, i.e., function-to-function, content-to-content and other combinations (function-to-content and content-to-function). These evaluations were conducted for a subset of words in both reference and system output selected by POS. For example, in function-to-function evaluation, we used alignments between source side function words and target side function words. Content-to-content and function-to-content were evaluated in the same way as the function-to-function evaluation. In the function-to-content evaluation, we used not only alignments between source side function word and target side content words,

but also used alignments between source side content words and target side function words. For discriminating function words from content words, we used the Stanford POS Tagger (Toutanova, Klein, D. Manning, and Singer 2003) for French and English sentences, and used KyTea (Neubig, Nakata, and Mori 2011a) for Japanese sentences. In French and English sentences, we treated a word with a label in a set {CC, CD, TO, DT, UH, EX, IN, LS, WDT, MD, WP, WP\$, WRB, PDT, PRP, PRP, RP, SYM, TO, UH} as a function word, and a set {VB, VBD, VBG, JJ, VBN, JJR, VBP, JJS, VBZ, NN, NNS, NNP, NNPS, RB, RBR, RBS} as a content word. In Japanese sentences, we treated a word with a label in a set { 代名詞 (pronoun), 接続詞 (conjunction), 感動詞 (interjection), 助動詞 (auxiliary verb), 助詞 (particle), 接頭辞 (prefix), 接尾辞 (postfix), 記号 (symbol), 補助記号 (diacritical mark) } as a function word, and a set { 名詞 (noun), 連体詞 (pre-noun adjectival), 副詞 (adverb), 形容詞 (adjective), 形状詞 (adjective noun), 動詞 (verb) } as a content word. When generating the final bidirectional word alignment, we used a grow-diag-final heuristic (Och and Ney 2003) for the Japanese-English task and an intersection heuristic for the French-English task, based on preliminary studies.

Table 3 summarizes our results against all word alignments in the references. The GDF in the tables denotes a heuristic extraction method grow-diag-final, and Filtered denotes the filtering method (Liang et al. 2006). The baseline method is symmetric constraint (Ganchev et al. 2010) shown in Table 3. The numbers in bold and in italics indicate the best score and the second best

Table 3 Results of word alignment evaluation

extraction	model	constraint	Hansard (French-English)				KFTT			
			precision	recall	F	AER	precision	recall	F	AER
Heuristic	HMM	<i>sym</i>	0.7135	0.8975	0.7950	<i>0.0665</i>	0.4215	<i>0.6004</i>	0.4953	0.5047
		<i>f2f</i>	0.7159	0.8975	0.7965	0.0662	0.4224	0.6023	0.4966	0.5034
		<i>c2c</i>	<i>0.7141</i>	0.8955	0.7946	0.0675	0.4190	0.5959	0.4920	0.5080
		<i>f2c</i>	0.7159	0.8950	<i>0.7955</i>	0.0678	0.4188	0.5970	0.4923	0.5077
	M4	<i>sym</i>	0.7021	0.8824	0.7820	0.0745	0.4620	0.5990	0.5216	0.4784
		<i>f2f</i>	0.7026	0.8816	0.7820	0.0748	0.4633	0.6001	0.5229	0.4771
		<i>c2c</i>	0.7025	0.8814	0.7819	0.0752	0.4605	0.5989	0.5207	0.4793
		<i>f2c</i>	0.7016	0.8796	0.7806	0.0752	0.4617	0.5994	0.5216	0.4784
Filtered	HMM	<i>sym</i>	0.5210	0.9475	0.6723	0.1077	0.5156	0.4281	0.4678	0.5322
		<i>f2f</i>	0.5218	0.9448	0.6723	0.1086	0.5109	0.4257	0.4645	0.5355
		<i>c2c</i>	0.5197	<i>0.9455</i>	0.6708	0.1109	0.5122	0.4242	0.4641	0.5359
		<i>f2c</i>	0.5209	<i>0.9455</i>	0.6718	0.1101	0.5121	0.4263	0.4653	0.5347
	M4	<i>sym</i>	0.5954	0.9165	0.7219	0.0872	0.7117	0.4542	0.5545	0.4455
		<i>f2f</i>	0.5911	0.9118	0.7172	0.0910	<i>0.7145</i>	0.4544	0.5555	0.4445
		<i>c2c</i>	0.5962	0.9160	0.7223	0.0877	0.7182	0.4535	<i>0.5560</i>	<i>0.4440</i>
		<i>f2c</i>	0.6001	0.9156	0.7250	0.0854	0.7143	0.4564	0.5570	0.4430

score, respectively. In terms of F-measure, it is clear that the $f2f$ is the most effective constraint for Hansard while we achieved the best score for the $f2c$ in KFTT, and both methods exceed the original posterior regularized model of Ganchev et al. (2010). In terms of AER, our proposed $f2c$ constraint has little gain against baseline symmetric constraint. On the other hand, our proposed constraints have no gain against baseline in Hansard corpus. Table 4, 5 and 6 summarizes our results by differentiating word alignment types, i.e., function-to-function, content-to-content and others, respectively. From these results, we can observe that our proposed constraint functions are effective for function-to-function and function-to-content word alignments in Hansard, and function-to-function and content-to-content word alignments in KFTT. These results match with our expectation that the baseline symmetric constraint is difficult to align the function word pairs, compared with our proposed constraints, as discussed in the introduction section. Figure 5–8 are recall-precision graph for each combination of corpus and models. We plotted the recall and precision by changing the threshold value of filtering method (Liang et al. 2006) on interval 0.01. From Figure 5, the precision value of our proposed $f2f$ constraint under HMM is better than that of the baseline symmetric constraint at low recall area. From Figure 6, in contrast to the HMM, the precision values of $f2f$ and $f2c$ constraints under IBM Model 4 are better than the baseline symmetric constraint at high recall area. However, from Figure 7 and 8, we observed unstable performance for Hansard task both on HMM and IBM Model 4 and it is hard to tell any

Table 4 Results of evaluation against function-to-function word alignments

extraction	model	constraint	Hansard (French-English)				KFTT			
			precision	recall	F	AER	precision	recall	F	AER
Heuristic	HMM	<i>sym</i>	0.7097	0.9706	0.8199	0.0311	0.4306	0.6256	0.5101	0.4899
		<i>f2f</i>	0.7416	0.9706	0.8408	0.0255	0.4365	0.6268	0.5146	0.4854
		<i>c2c</i>	<i>0.7283</i>	<i>0.9853</i>	<i>0.8375</i>	0.0312	0.4305	0.6316	0.5120	0.4880
		<i>f2c</i>	0.7097	0.9706	0.8199	0.0373	0.4345	<i>0.6365</i>	0.5164	0.4836
	M4	<i>sym</i>	0.7021	0.9706	0.8148	<i>0.0309</i>	0.4824	0.6280	0.5456	0.4544
		<i>f2f</i>	0.6598	0.9412	0.7758	0.0545	0.4912	0.6389	0.5554	0.4446
		<i>c2c</i>	0.6771	0.9559	0.7927	0.0366	0.4861	0.6341	0.5503	0.4497
		<i>f2c</i>	0.6947	0.9706	0.8098	0.0307	0.4967	<i>0.6365</i>	0.5580	0.4420
Filtered	HMM	<i>sym</i>	0.5191	1.000	0.6834	0.0905	0.5379	0.4626	0.4974	0.5026
		<i>f2f</i>	0.5271	1.000	0.6904	0.0964	0.5428	0.4517	0.4931	0.5069
		<i>c2c</i>	0.5113	1.000	0.6766	0.0945	0.5343	0.4517	0.4895	0.5105
		<i>f2c</i>	0.5191	1.000	0.6834	0.1005	0.5381	0.4517	0.4911	0.5089
	M4	<i>sym</i>	0.6036	<i>0.9853</i>	0.7486	0.0726	<i>0.7330</i>	0.4807	0.5806	0.4194
		<i>f2f</i>	0.5641	0.9706	0.7135	0.0973	0.7418	0.4891	0.5895	0.4105
		<i>c2c</i>	0.5690	0.9706	0.7174	0.0761	0.7200	0.4783	0.5747	0.4253
		<i>f2c</i>	0.5690	0.9706	0.7174	0.0870	0.7371	0.4843	<i>0.5845</i>	<i>0.4155</i>

Table 5 Results of evaluation against content-to-content word alignments

extraction	model	constraint	Hansard (French-English)				KFTT			
			precision	recall	F	AER	precision	recall	F	AER
Heuristic	HMM	<i>sym</i>	0.6835	0.9087	0.7802	0.0753	0.4600	<i>0.6660</i>	0.5442	0.4558
		<i>f2f</i>	0.6957	0.9231	0.7934	0.0671	0.4630	0.6674	0.5467	0.4533
		<i>c2c</i>	0.6841	0.9111	0.7814	0.0722	0.4612	0.6632	0.5441	0.4559
		<i>f2c</i>	0.6848	0.9087	0.7810	0.0713	0.4623	0.6632	0.5448	0.4552
	M4	<i>sym</i>	<i>0.6882</i>	0.8966	0.7787	<i>0.0647</i>	0.5040	0.6569	0.5704	0.4296
		<i>f2f</i>	0.6424	0.9038	<i>0.7842</i>	0.0636	0.5117	0.6569	0.5753	0.4247
		<i>c2c</i>	0.6850	0.8990	0.7775	0.0665	0.5029	0.6604	0.5710	0.4290
		<i>f2c</i>	0.6788	0.8990	0.7735	0.0651	0.5094	0.6597	0.5749	0.4251
Filtered	HMM	<i>sym</i>	0.4902	<i>0.9591</i>	0.6488	0.1008	0.5469	0.4310	0.4821	0.5179
		<i>f2f</i>	0.4920	0.9567	0.6498	0.1037	0.5496	0.4289	0.4818	0.5182
		<i>c2c</i>	0.4902	0.9615	0.6494	0.1063	0.5437	0.4247	0.4769	0.5231
		<i>f2c</i>	0.4866	0.9567	0.6451	0.1045	0.5464	0.4268	0.4792	0.5208
	M4	<i>sym</i>	0.5754	0.9351	0.7125	0.0815	<i>0.7198</i>	0.5105	0.5973	0.4027
		<i>f2f</i>	0.5646	0.9351	0.7041	0.0842	0.7164	0.4986	0.5880	0.4120
		<i>c2c</i>	0.5631	0.9327	0.7023	0.0869	0.7256	0.4979	0.5906	0.4094
		<i>f2c</i>	0.5669	0.9375	0.7065	0.0815	0.7137	0.5042	<i>0.5909</i>	<i>0.4091</i>

Table 6 Results of evaluation against function-to-content and content-to-function word alignments

extraction	model	constraint	Hansard (French-English)				KFTT			
			precision	recall	F	AER	precision	recall	F	AER
Heuristic	HMM	<i>sym</i>	0.7173	0.8948	<i>0.7963</i>	0.0661	0.4544	0.6324	0.5288	0.4712
		<i>f2f</i>	<i>0.7179</i>	0.8931	0.7960	<i>0.0668</i>	0.4552	<i>0.6315</i>	0.5290	0.4710
		<i>c2c</i>	0.7175	0.8920	0.7953	0.0676	0.4551	<i>0.6336</i>	0.5297	0.4703
		<i>f2c</i>	0.7200	0.8920	0.7968	0.0680	0.4522	0.6285	0.5259	0.4741
	M4	<i>sym</i>	0.7038	0.8790	0.7817	0.0766	0.4985	0.6298	0.5565	0.4435
		<i>f2f</i>	0.7048	0.8779	0.7819	0.0766	0.4949	0.6302	0.5544	0.4456
		<i>c2c</i>	0.7052	0.8779	0.7822	0.0771	0.4929	0.6242	0.5508	0.4492
		<i>f2c</i>	0.7046	0.8756	0.7808	0.0773	0.4922	0.6268	0.5514	0.4486
Filtered	HMM	<i>sym</i>	0.5249	0.9451	0.6750	0.1089	0.5474	0.4223	0.4768	0.5232
		<i>f2f</i>	0.5255	0.9423	0.6747	0.1095	0.5390	0.4197	0.4719	0.5281
		<i>c2c</i>	0.5237	0.9426	0.6733	0.1117	0.5455	0.4184	0.4736	0.5264
		<i>f2c</i>	0.5254	<i>0.9432</i>	0.6749	0.1110	0.5462	0.4223	0.4763	0.5237
	M4	<i>sym</i>	0.5977	0.9131	0.7225	0.0882	0.7309	0.4748	<i>0.5757</i>	<i>0.4243</i>
		<i>f2f</i>	0.5951	0.9080	0.7189	0.0917	<i>0.7322</i>	0.4684	0.5713	0.4287
		<i>c2c</i>	0.6010	0.9131	0.7249	0.0880	0.7349	0.4701	0.5734	0.4266
		<i>f2c</i>	0.6050	0.9119	0.7274	0.0858	0.7413	0.4761	0.5798	0.4202

significant difference with each other. These results may indicate that our proposed constraint is suitable for grammatically different language pairs, such as Japanese and English, under word alignment task. We will provide more details in Section 6.

Fig. 5 Recall-Precision graph of HMM in KFTT Japanese-English

Fig. 6 Recall-Precision graph of IBM Model 4 in KFTT Japanese-English

Fig. 7 Recall-Precision graph of HMM in Hansard French-English

Fig. 8 Recall-Precision graph of IBM Model 4 in Hansard French-English

5.3 Translation Evaluation

Next, we performed a translation evaluation, measured by BLEU (Papineni, Roukos, Ward, and Zhu 2002). We used the grow-diag-final and filtering method (Liang et al. 2006) for creating phrase tables. The threshold for the filtering factor was set to 0.1. From the English side of the training data, we trained a 5-gram language models with SRILM (Stolcke 2002). “Moses” toolkit was used as a phrase based decoder (Koehn, Hoang, Birch, Callison-Burch, Federico, Bertoldi, Cowan, Shen, Moran, Zens, Dyer, Bojar, Constantin, and Herbst 2007) and the model parameters were tuned by k-best MIRA (Cherry and Foster 2012). We set the distortion-limit as -1 (infinite). We used other parameters with Moses default values. In order to avoid tuning instability, we evaluated the average of five runs (Hopkins and May 2011).

The results are summarized in Table 7. The numbers in bold style are not significantly different from the best results (p-value < 0.05). We used statistical bootstrapping method (Koehn 2004) for significance test. Our proposed constraints achieved significant improvement against baseline symmetric constraint in NTCIR10 task and europarl-v7 Spanish-English, but observed no gain in the other language pairs of europarl-v7. These results clearly show that our proposed constraints are effective on grammatically different language pairs as observed in the word alignment evaluation. The BLEU score of grow-diag-final is better than that of filtered method in

Table 7 Results of translation evaluation

extraction	model	constraint	europarl-v7			NTCIR10
			de-en	es-en	fr-en	ja-en
GDF	HMM	<i>sym</i>	26.59	31.78	31.44	29.21
		<i>f2f</i>	26.80	31.83	31.57	29.28
		<i>c2c</i>	26.67	31.64	31.53	29.23
		<i>f2c</i>	26.80	31.76	31.48	29.21
	M4	<i>sym</i>	27.16	31.86	31.31	29.59
		<i>f2f</i>	27.10	31.78	31.50	29.58
		<i>c2c</i>	27.05	31.76	31.25	29.68
		<i>f2c</i>	27.13	31.75	31.49	29.79
Filtered	HMM	<i>sym</i>	26.97	31.86	31.52	29.00
		<i>f2f</i>	26.96	31.78	31.53	29.01
		<i>c2c</i>	26.94	31.90	31.56	29.02
		<i>f2c</i>	26.84	31.99	31.54	29.02
	M4	<i>sym</i>	26.80	31.85	31.31	29.42
		<i>f2f</i>	27.01	31.77	31.38	29.40
		<i>c2c</i>	26.91	31.72	31.34	29.32
		<i>f2c</i>	26.88	31.78	31.34	29.40

NTCIR10, whereas the other language pairs have no difference. The BLEU scores of IBM Model4 and HMM are not different from each other except for NTCIR10. From these results, we can conclude that the IBM Model 4 combined with our proposed constraints under grow-diag-final is the most robust method to perform word alignment for grammatically different language pairs indicated by the end-to-end evaluation.

6 Analysis

We found that our proposed methods were superior to the strong baseline in terms of recall-precision and AER for KFTT in the word alignment evaluations. However, in Hansard Fr-En, proposed constraints have little advantage against the baseline. Our $f2f$, $c2c$ and $f2c$ constraints can perform accurate word alignment either by rewarding or penalising certain pairings of function and content words. In a grammatically different language pairs, such as Japanese and English, function words and content words are sometimes misaligned due to the one-to-many limitation of generative models. As a result, the proposed constraints may work better on KFTT Japanese-English corpus. Figure 9 and 10 show examples of word alignment produced by IBM Model 4, in

Fig. 9 An example of a sentence pair in KFTT aligned by IBM Model4 with and without $f2f$ constraints

Fig. 10 Another example of a sentence pair in KFTT aligned by IBM Model4 with and without $f2f$ constraints

which circles and squares indicate word alignment with and without $f2f$ constraints, respectively. The circles and squares painted in black represent the difference between with and without $f2f$ constraints. In Figure 9, Japanese words “*muromachi*” and “*dai*” should be correctly aligned to “NULL” and an English word “the”. The baseline symmetric constraint misaligned “*muromachi*” and “the”. On the other hand, the $f2f$ constraint correctly aligned “*dai*” and “the” by rewarding the pair of highly frequent words. In Figure 10, Japanese words “*ninmei*” and “*rare*” should be correctly aligned to English words “succeeded” and “were”. Although the Japanese word “*rare*” and the English word “were” were correctly aligned under the $f2f$ constraint, misalignment was observed for the Japanese word “*ninmei*” and the English word “post”. Since both “*rare*” and “were” are function words and used for a passive voice, our proposed constraint can easily align the two words. However, at the alignment of “*ninmei sare*” and “succeeded”, “*ninmei sare*” is a passive voice and “succeeded” is an active voice. Because of this, our proposed constraint cannot easily align these words. In addition to this difficulty, as the Japanese phrase “*kounin no kanrei niwa yoshimasa ga ninmei sare*” are translated into English phrase “yoshimasa succeeded to the post of kanrei” by expert translators, our $f2f$ constraint could not correct the missed alignment. In general, it is difficult to perform correct word alignment for those phraseological expressions especially when models are strongly restricted to one-to-many correspondence.

7 Conclusion

In this paper, we proposed new constraint functions under the posterior regularization framework. Our constraint functions introduce fine-grained agreement constraint considering the frequency of words, assuming that the high frequency words correspond to function words whereas the less frequent words may be treated as content words, based on the previous work (Setiawan et al. 2007). Experiments on word alignment tasks showed better alignment qualities measured by F-measure and AER on KFTT. We also observed large gain in BLEU, 0.2 on average, when compared with the previous posterior regularization method under NTCIR10 task. From the analysis of our constraint, we can conclude that considering phrase structure is necessary for better alignment and translation quality. As our future work, we will investigate the phrase structure constraint, in reference to the previous many-to-many word alignment research, such as ITG based (Neubig et al. 2011b) or syntactic tree based (Nakazawa and Kurohashi 2012) models. Our proposed constraints are based on feature functions, and these functions are used in a combinatorial manner on various NLP tasks. However, we used a single feature function for generating word alignments. We will also investigate the combinatory use of our proposed constraints for better alignment and translation quality.

Acknowledgement

We would like to thank the reviewers for their detailed comments and suggestions, which are helpful to improve our manuscripts.

Reference

- Bisazza, A. and Federico, M. (2012). “Cutting the Long Tail: Hybrid Language Models for Translation Style Adaptation.” In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 439–448, Avignon, France. Association for Computational Linguistics.
- Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). “The Mathematics of Statistical Machine Translation: Parameter Estimation.” *Computational Linguistics*, **19** (2), pp. 263–311.
- Chang, Y.-W., Rush, A. M., DeNero, J., and Collins, M. (2014). “A Constrained Viterbi Relaxation for Bidirectional Word Alignment.” In *Proceedings of the 52nd Annual Meeting*

- of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1481–1490, Baltimore, Maryland. Association for Computational Linguistics.
- Cherry, C. and Foster, G. (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*, pp. 427–436, Montréal, Canada. Association for Computational Linguistics.
- DeNero, J., Bouchard-Côté, A., and Klein, D. (2008). “Sampling Alignment Structure under a Bayesian Translation Model.” In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 314–323. Association for Computational Linguistics.
- DeNero, J. and Macherey, K. (2011). “Model-Based Aligner Combination Using Dual Decomposition.” In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 420–429, Portland, Oregon, USA. Association for Computational Linguistics.
- Fraser, A. and Marcu, D. (2007). “Measuring Word Alignment Quality for Statistical Machine Translation.” *Computational Linguistics*, **33** (3), pp. 293–303.
- Ganchev, K., Graca, J., Gillenwater, J., and Taskar, B. (2010). “Posterior Regularization for Structured Latent Variable Models.” *The Journal of Machine Learning Research*, **11**, pp. 2001–2049.
- Goto, I., Chow, K. P., Lu, B., Sumita, E., and Tsou, B. K. (2013). “Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop.” In *Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-10*, pp. 260–286.
- Hopkins, M. and May, J. (2011). “Tuning as Ranking.” In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1352–1362, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Kamigaito, H., Watanabe, T., Takamura, H., and Okumura, M. (2014). “Unsupervised Word Alignment Using Frequency Constraint in Posterior Regularized EM.” In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 153–158, Doha, Qatar. Association for Computational Linguistics.
- Koehn, P. (2004). “Statistical Significance Tests for Machine Translation Evaluation.” In Lin, D. and Wu, D. (Eds.), *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen,

- W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (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*, pp. 177–180, Prague, Czech Republic. Association for Computational Linguistics.
- Koehn, P., Och, F. J., and Marcu, D. (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*, pp. 48–54. Association for Computational Linguistics.
- Liang, P., Taskar, B., and Klein, D. (2006). “Alignment by Agreement.” In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pp. 104–111, New York City, USA. Association for Computational Linguistics.
- Matusov, E., Zens, R., and Ney, H. (2004). “Symmetric Word Alignments for Statistical Machine Translation.” In *Proceedings of COLING 2004, the 20th International Conference on Computational Linguistics*, pp. 219–225, Geneva, Switzerland. COLING.
- Nakazawa, T. and Kurohashi, S. (2012). “Alignment by Bilingual Generation and Monolingual Derivation.” In *Proceedings of COLING 2012, the 24th International Conference on Computational Linguistics*, pp. 1963–1978. The COLING 2012 Organizing Committee.
- Neubig, G. (2011). “The Kyoto Free Translation Task.” <http://www.phontron.com/kfft>.
- Neubig, G., Nakata, Y., and Mori, S. (2011a). “Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis.” In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 529–533. Association for Computational Linguistics.
- Neubig, G., Watanabe, T., Sumita, E., Mori, S., and Kawahara, T. (2011b). “An Unsupervised Model for Joint Phrase Alignment and Extraction.” In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 632–641. Association for Computational Linguistics.
- Och, F. J. and Ney, H. (2003). “A Systematic Comparison of Various Statistical Alignment Models.” *Computational Linguistics*, **29** (1), pp. 19–51.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). “Bleu: A Method for Automatic Evaluation of Machine Translation.” In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Setiawan, H., Kan, M.-Y., and Li, H. (2007). “Ordering Phrases with Function Words.” In

- Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 712–719, Prague, Czech Republic. Association for Computational Linguistics.
- Stolcke, A. (2002). “SRILM—An Extensible Language Modeling Toolkit.” In *Proceedings of the International Conference on Spoken Language Processing*, pp. 257–286.
- Toutanova, K., Klein, D., D. Manning, C., and Singer, Y. (2003). “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network.” In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 173–180.
- Vogel, S., Ney, H., and Tillmann, C. (1996). “HMM-based Word Alignment in Statistical Translation.” In *Proceedings of the 16th conference on Computational Linguistics-Volume 2*, pp. 836–841. Association for Computational Linguistics.

Hidetaka Kamigaito: Hidetaka Kamigaito recieved B.E. and M.E. from Tokyo Institute of Technology in 2012 and 2014. He is currently a Ph.D. student of Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology. His current research interests are statistical machine translation and constituent parsing.

Taro Watanabe: Taro Watanabe received the B.E. and M.E. degrees in informaiton science from Kyoto Univ., Kyoto, Japan in 1994 and 1997, respectively, and obtained the Master of Science degree in language and information technologies from the School of Computer Science, Carnegie Mellon University in 2000. In 2004, he received the Ph.D. in informatics from Kyoto Univ., Kyoto, Japan. After working as a researcher at ATR, NTT and NICT, Dr. Watanabe is a software engineer at Google, Inc. His research interests include natural language processing, machine learning and statistical machine translation.

Hiroya Takamura: Hiroya Takamura received B.E. and M.E. from the University of Tokyo in 1997 and 2000 respectively (in 1999 he was a research student at Technische Universitaet von Wien). He received Dr. Eng. from Nara Institute of Science and Technology in 2003. He was an assistant professor at Tokyo Institute of Technology from 2003 to 2010. He is currently an associate professor at Tokyo Institute of Technology. His current research interest is computational linguistics. He is a member of the Information Processing

Society of Japan and the Association for Computational Linguistics.

Manabu Okumura: Manabu Okumura was born in 1962. He received B.E., M.E. and Dr. Eng. from Tokyo Institute of Technology in 1984, 1986 and 1989 respectively. He was an assistant at the Department of Computer Science, Tokyo Institute of Technology from 1989 to 1992, and an associate professor at the School of Information Science, Japan Advanced Institute of Science and Technology from 1992 to 2000. He is currently a professor at Institute of Innovative Research, Tokyo Institute of Technology. His current research interests include natural language processing, especially text summarization, computer assisted language learning, sentiment analysis, and text data mining.

Eiichiro Sumita: Dr. Eiichiro Sumita received his Ph.D. in Engineering from Kyoto University in 1999, and a Master's and Bachelor's in Computer Science from the University of Electro-Communications in 1982 and 1980, respectively. He is now in the National Institute of Information and Communication Technology (NICT) , its fellow and the associate director-general of Advanced Speech Translation Research and Development Promotion Center (ASTREC). His research interests cover Machine Translation and e-Learning. He is a co-recipient of the Maejima Hisoka Prize in 2013, the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, Prizes for Science and Technology in 2010, and the AAMT Nagao Award in 2007.

(Received September 11, 2015)

(Revised December 3, 2015)

(Rerevised March 14, 2016)

(Accepted April 8, 2016)