Unsupervised Word Alignment Using Frequency Constraint in Posterior Regularized EM

Hidetaka Kamigaito[1,2], **Taro Watanabe**[2], **Hiroya Takamura**[1], **Manabu Okumura**[1]
[1]Tokyo Institute of Technology, Precision and Intelligence Laboratory
4259 Nagatsuta-cho Midori-ku Yokohama, Japan
[2]National Institute of Information and Communication Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, Japan

Abstract

Generative word alignment models, such as IBM Models, are restricted to one-to-many alignment, and cannot explicitly represent many-to-many relationships in a bilingual text. The problem is partially solved either by introducing heuristics or by agreement constraints such that two directional word alignments agree with each other. In this paper, we focus on the posterior regularization framework (Ganchev et al., 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. Experimental results on French-to-English and Japanese-to-English alignment tasks show statistically significant gains over the previous posterior regularization baseline. We also observed gains in Japanese-to-English translation tasks, which prove the effectiveness of our methods under grammatically different language pairs.

1 Introduction

Word alignment is an important component in statistical machine translation (SMT). For instance phrase-based SMT (Koehn et al., 2003) is based on the concept of phrase pairs that are automatically extracted from bilingual data and rely on 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.

The Generative word alignment models, such as the IBM Models (Brown et al., 1993) and HMM (Vogel et al., 1996), are popular methods for automatically aligning bilingual texts, but 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 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. 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 et al. (2014) also enforce agreement during decoding.

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.

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 differentiate between content words and function words by frequency in bilingual data, following Setiawan et al. (2007).

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

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 153–158,
October 25-29, 2014, Doha, Qatar.

in BLEU scores in the NTCIR10.

2 Statistical word alignment with posterior regularization framework

Given a bilingual sentence $\boldsymbol{x} = (\boldsymbol{x}^s, \boldsymbol{x}^t)$ where $\boldsymbol{x}^s$ and $\boldsymbol{x}^t$ denote a source and target sentence, respectively, the bilingual sentence is aligned by a many-to-many alignment of $\boldsymbol{y}$. We represent posterior probabilities from two directional word alignment models as $\overrightarrow{p}_\theta(\overrightarrow{\boldsymbol{y}}|\boldsymbol{x})$ and $\overleftarrow{p}_\theta(\overleftarrow{\boldsymbol{y}}|\boldsymbol{x})$ with each arrow indicating a particular direction, and use θ to denote the parameters of the models. For instance, $\overrightarrow{y}$ is a subset of $\boldsymbol{y}$ for the alignment from $\boldsymbol{x}^s$ to $\boldsymbol{x}^t$ under the model of $p(\boldsymbol{x}^t, \overrightarrow{\boldsymbol{y}}|\boldsymbol{x}^s)$. In the case of IBM Model 1, the model is represented as follows:

$$p(\boldsymbol{x}^t, \overrightarrow{\boldsymbol{y}}|\boldsymbol{x}^s) = \prod_i \frac{1}{|\boldsymbol{x}^s|+1} p_t(x_i^t|\boldsymbol{x}^s_{\overrightarrow{y}_i}). \quad (1)$$

where we define the index of $\boldsymbol{x}^t$, $\boldsymbol{x}^s$ as $i, j (1 \leq i \leq |\boldsymbol{x}^t|, 1 \leq j \leq |\boldsymbol{x}^s|)$ and the posterior probability for the word pair (x_i^t, x_j^s) is defined as follows:

$$\overrightarrow{p}(i, j|\boldsymbol{x}) = \frac{p_t(x_i^t|x_j^s)}{\sum_{j'} p_t(x_i^t|x_{j'}^s)}. \quad (2)$$

Herein, we assume that the posterior probability for wrong directional alignment is zero (i.e., $\overrightarrow{p}(\overleftarrow{\boldsymbol{y}}|\boldsymbol{x}) = 0$).[1] Given the two directional models, Ganchev et al. defined a symmetric feature for each target/source position pair, i, j as follows:

$$\phi_{i,j}(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} +1 & (\overrightarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overrightarrow{y}_i = j), \\ -1 & (\overleftarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overleftarrow{y}_j = i), \\ 0 & otherwise. \end{cases} \quad (3)$$

The feature assigns 1 for the subset of word alignment for $\overrightarrow{\boldsymbol{y}}$, but assigns -1 for $\overleftarrow{\boldsymbol{y}}$. 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. Ganchev et al. defined a joint model that combines two directional models using arithmetic means:

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

Under the posterior regularization framework, we instead use q that is derived by maximizing the following posterior probability parametrized by $\boldsymbol{\lambda}$ for each bilingual data $\boldsymbol{x}$ as follows (Ganchev et al., 2010):

$$q_{\boldsymbol{\lambda}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\overrightarrow{p}_\theta(\overrightarrow{\boldsymbol{y}}|\boldsymbol{x}) + \overleftarrow{p}_\theta(\overleftarrow{\boldsymbol{y}}|\boldsymbol{x})}{2} \cdot \frac{exp\{-\boldsymbol{\lambda}\cdot\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\}}{Z} \quad (5)$$

$$= \frac{\overrightarrow{q}(\overrightarrow{\boldsymbol{y}}|\boldsymbol{x})\frac{Z_{\overrightarrow{q}}}{\overrightarrow{p}_\theta(\boldsymbol{x})} + \overleftarrow{q}(\overleftarrow{\boldsymbol{y}}|\boldsymbol{x})\frac{Z_{\overleftarrow{q}}}{\overleftarrow{p}_\theta(\boldsymbol{x})}}{2Z},$$

$$Z = \frac{1}{2}\left(\frac{Z_{\overrightarrow{q}}}{\overrightarrow{p}_\theta} + \frac{Z_{\overleftarrow{q}}}{\overleftarrow{p_\theta}}\right),$$

$$\overrightarrow{q}(\overrightarrow{\boldsymbol{y}}|\boldsymbol{x}) = \frac{1}{Z_{\overrightarrow{q}}}\overrightarrow{p}_\theta(\overrightarrow{\boldsymbol{y}}, \boldsymbol{x}) exp\{-\boldsymbol{\lambda}\cdot\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\},$$

$$Z_{\overrightarrow{q}} = \sum_{\overrightarrow{y}} \overrightarrow{p}_\theta(\overrightarrow{\boldsymbol{y}}, \boldsymbol{x}) exp\{-\boldsymbol{\lambda}\cdot\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\},$$

$$\overleftarrow{q}(\overleftarrow{\boldsymbol{y}}|\boldsymbol{x}) = \frac{1}{Z_{\overleftarrow{q}}}\overleftarrow{p}_\theta(\overleftarrow{\boldsymbol{y}}, \boldsymbol{x}) exp\{-\boldsymbol{\lambda}\cdot\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\},$$

$$Z_{\overleftarrow{q}} = \sum_{\overleftarrow{y}} \overleftarrow{p}_\theta(\overleftarrow{\boldsymbol{y}}, \boldsymbol{x}) exp\{-\boldsymbol{\lambda}\cdot\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\},$$

such that $\mathbb{E}_{q_{\boldsymbol{\lambda}}}[\phi_{i,j}(\boldsymbol{x}, \boldsymbol{y})] = 0$. In the E-step of EM-algorithm, we employ $q_{\boldsymbol{\lambda}}$ instead of p_θ to accumulate fractional counts for its use in the M-step. $\boldsymbol{\lambda}$ is efficiently estimated by the gradient ascent for each bilingual sentence $\boldsymbol{x}$. Note that posterior regularization is performed during parameter estimation, and not during testing.

3 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 aligned one-to-one, but function words are not always aligned together. In addition, Japanese is a pro-drop language which can easily violate the symmetric constraint when proper nouns in the English side have to be aligned with a "null" word. In addition, low frequency words may cause unreliable estimates for adjusting the weighing parameters $\boldsymbol{\lambda}$.

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.

3.1 Mismatching constraint

First, we propose a mismatching constraint that penalizes word alignment between content words and function words by decreasing the corresponding posterior probabilities.

[1] No alignment is represented by alignment into a special token "null".

The constraint is represented as *f2c* (*function to content*) constraint:

$$\phi^{\text{f2c}}_{i,j}(\boldsymbol{x},\boldsymbol{y}) = \tag{6}$$
$$\begin{cases} +1 & (\overrightarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overrightarrow{y}_i = j) \cap ((x^t_i \in \mathcal{C}^t \cap x^s_j \in \mathcal{F}^s) \\ & \cup (x^t_i \in \mathcal{F}^t \cap x^s_j \in \mathcal{C}^s)) \cap (\delta_{i,j}(\boldsymbol{x},\boldsymbol{y}) > 0), \\ 0 & (\overleftarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overleftarrow{y}_j = i) \cap ((x^t_i \in \mathcal{C}^t \cap x^s_j \in \mathcal{F}^s) \\ & \cup (x^t_i \in \mathcal{F}^t \cap x^s_j \in \mathcal{C}^s)) \cap (\delta_{i,j}(\boldsymbol{x},\boldsymbol{y}) > 0), \\ 0 & (\overrightarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overrightarrow{y}_i = j) \cap ((x^t_i \in \mathcal{C}^t \cap x^s_j \in \mathcal{F}^s) \\ & \cup (x^t_i \in \mathcal{F}^t \cap x^s_j \in \mathcal{C}^s)) \cap (\delta_{i,j}(\boldsymbol{x},\boldsymbol{y}) < 0), \\ -1 & (\overleftarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overleftarrow{y}_j = i) \cap ((x^t_i \in \mathcal{C}^t \cap x^s_j \in \mathcal{F}^s) \\ & \cup (x^t_i \in \mathcal{F}^t \cap x^s_j \in \mathcal{C}^s)) \cap (\delta_{i,j}(\boldsymbol{x},\boldsymbol{y}) < 0). \end{cases}$$

where $\delta_{i,j}(\boldsymbol{x},\boldsymbol{y}) = \overrightarrow{p}_\theta(i,j|\boldsymbol{x}) - \overleftarrow{p}_\theta(i,j|\boldsymbol{x})$ is the difference in the posterior probabilities between the source-to-target and the target-to-source alignment. $\mathcal{C}^s$ 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 (i,j), the constraint function returns a non-zero 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.

Note that when this constraint is not fired, we fall back to the constraint function in Equation (3) for each word pair.

3.2 Matching constraint

In contrast to the mismatching constraint, our second constraint function rewards alignment for *function to function* word matching, namely *f2f*. The f2f constraint function is defined as follows:

$$\phi^{\text{f2f}}_{i,j}(\boldsymbol{x},\boldsymbol{y}) = \tag{7}$$
$$\begin{cases} +1 & (\overrightarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overrightarrow{y}_i = j) \cap \\ & (x^t_i \in \mathcal{F}^t \cap x^s_j \in \mathcal{F}^s) \cap (\delta_{i,j}(\boldsymbol{x},\boldsymbol{y}) > 0), \\ 0 & (\overleftarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overleftarrow{y}_j = i) \cap \\ & (x^t_i \in \mathcal{F}^t \cap x^s_j \in \mathcal{F}^s) \cap (\delta_{i,j}(\boldsymbol{x},\boldsymbol{y}) > 0), \\ 0 & (\overrightarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overrightarrow{y}_i = j) \cap \\ & (x^t_i \in \mathcal{F}^t \cap x^s_j \in \mathcal{F}^s) \cap (\delta_{i,j}(\boldsymbol{x},\boldsymbol{y}) < 0), \\ -1 & (\overleftarrow{\boldsymbol{y}} \subset \boldsymbol{y}) \cap (\overleftarrow{y}_j = i) \cap \\ & (x^t_i \in \mathcal{F}^t \cap x^s_j \in \mathcal{F}^s) \cap (\delta_{i,j}(\boldsymbol{x},\boldsymbol{y}) < 0). \end{cases}$$

This constraint function returns a non-zero value for a word pair (i,j) when they are function words. As a result, the pair of function words are encouraged to agree with each other, but not other pairs. The *content to content* word matching function *c2c* can be defined similarly by replacing F^s and F^t by C^s and C^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. As noted in the mismatch function, when no constraint is fired, we fall back to Eq (3) for each word pair.

4 Experiment

4.1 Experimental Setup

The data sets used in our experiments are the French-English Hansard Corpus, and two data sets for Japanese-English tasks: the Kyoto free translation task (KFTT) and NTCIR10. The Hansard Corpus consists of parallel texts drawn from official records of the proceedings of the Canadian Parliament. The KFTT (Neubig, 2011) is derived from Japanese Wikipedia articles related to Kyoto, which is professionally translated into English. NTCIR10 comes from patent data employed in a machine translation shared task (Goto et al., 2013). The statistics of these data are presented in Table 1.

Sentences of over 40 words on both source and target sides are removed for training alignment models. We used a word alignment toolkit cicada [2] for training the IBM Model 4 with our proposed methods. Training is bootstrapped from IBM Model 1, followed by HMM and IBM Model 4. When generating the final bidirectional word alignment, we use a grow-diag-final heuristic for the Japanese-English tasks and an intersection heuristic in the French-English task, judged by preliminary studies.

Following Bisazza and Federico (2012), we automatically decide the threshold for word frequency to discriminate between content words and function words. Specifically, the threshold is determined by the ratio of highly frequent words. The threshold th is the maximum frequency that satisfies the following equation:

$$\frac{\sum_{w \in (freq(w) > th)} freq(w)}{\sum_{w \in all} freq(w)} > r. \tag{8}$$

Here, we empirically set $r = 0.5$ by preliminary studies. This method is based on the intuition that content words and function words exist in a document at a constant rate.

4.2 Word alignment evaluation

We measure the impact of our proposed methods on the quality of word alignment measured

[2] https://github.com/tarowatanabe/cicada

Table 1: The statistics of the data sets

			hansard		kftt		NTCIR10	
			French	English	Japanese	English	Japanese	English
train		sentence	1.13M		329.88K		2.02M	
		word	23.3M	19.8M	6.08M	5.91M	53.4M	49.4M
		vocabulary	78.1K	57.3K	114K	138K	114K	183K
dev		sentence			1.17K		2K	
		word			26.8K	24.3K	73K	67.3K
		vocabulary			4.51K	4.78K	4.38K	5.04K
test	WA	sentence	447		582			
		word	7.76K	7.02K	14.4K	12.6K		
		vocabulary	1,92K	1.69K	2.57K	2.65K		
	TR	sentence			1.16K		8.6K	
		word			28.5K	26.7K	334K	310K
		vocabulary			4.91K	4.57K	10.4K	12.7K

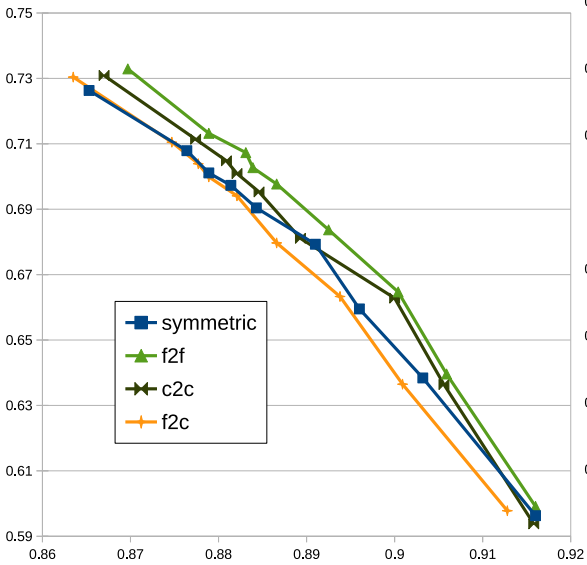

Figure 1: Precision Recall graph in Hansard French-English

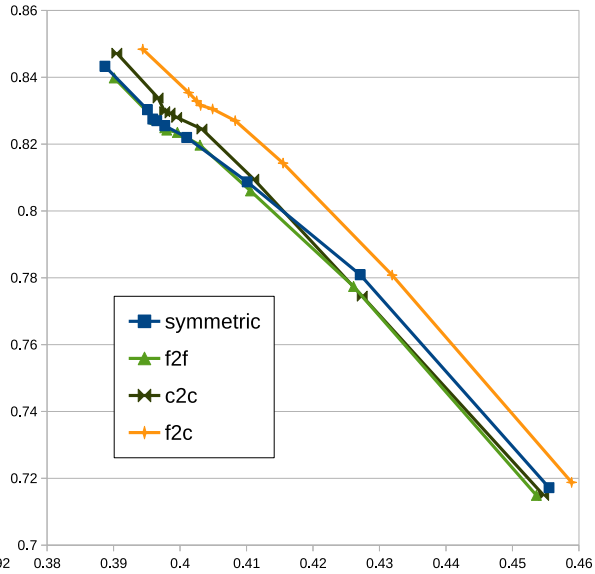

Figure 2: Precision Recall graph in KFTT

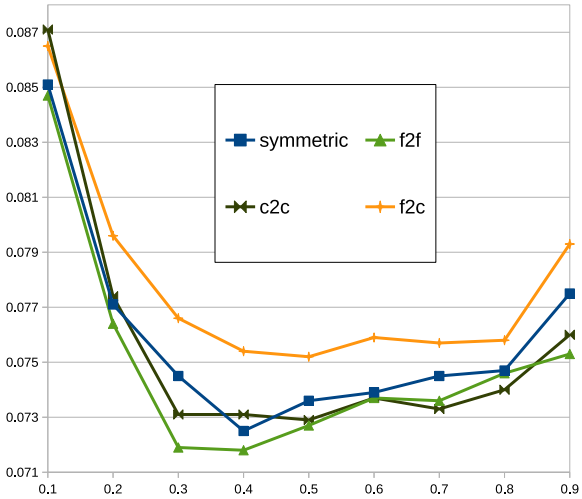

Figure 3: AER in Hansard French-English

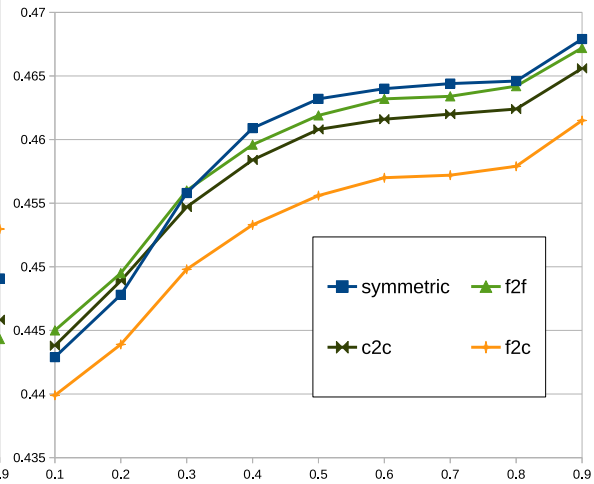

Figure 4: AER in KFTT

Table 2: Results of word alignment evaluation with the heuristics-based method (GDF)

	KFTT				Hansard (French-English)			
method	precision	recall	AER	F	precision	recall	AER	F
symmetric	0.4595	0.5942	48.18	0.5182	0.7029	0.8816	*7.29*	0.7822
f2f	**0.4633**	*0.5997*	**47.73**	**0.5227**	**0.7042**	*0.8851*	*7.29*	*0.7844*
c2c	0.4606	0.5964	48.02	0.5198	0.7001	0.8816	7.34	0.7804
f2c	*0.4630*	**0.5998**	*47.74*	*0.5226*	*0.7037*	**0.8871**	**7.10**	**0.7848**

by AER and F-measure (Och and Ney, 2003). Since there exists no distinction for sure-possible alignments in the KFTT data, we use only sure alignment for our evaluation, both for the French-English and the Japanese-English tasks. Table 2 summarizes our results.

The baseline method is symmetric constraint (Ganchev et al., 2010) shown in Table 2. The numbers in bold and in italics indicate the best score and the second best score, respectively. The differences between f2f,f2c and baseline in KFTT are statistically significant at $p < 0.05$ using the sign-test, but in hansard corpus, there exist no significant differences between the baseline and the proposed methods. In terms of F-measure, it is clear that the f2f method is the most effective method in KFTT, and both f2f and f2c methods exceed the original posterior regularized model of Ganchev et al. (2010).

We also compared these methods with filtering methods (Liang et al., 2006), in addition to heuristic methods. We plot precision/recall curves and AER by varying the threshold between 0.1 and 0.9 with 0.1 increments. From Figures, it can be seen that our proposed methods are superior to the baseline in terms of both precision-recall and AER.

4.3 Translation evaluation

Next, we performed a translation evaluation, measured by BLEU (Papineni et al., 2002). We compared 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 which was the best setting in the word alignment experiment in section 4.2 under KFTT. From the English side of the training data, we trained a word using the 5-gram model with SRILM (Stolcke and others, 2002). "Moses" toolkit was used as a decoder (Koehn et al., 2007) and the model parameters were tuned by k-best MIRA (Cherry and Foster, 2012). In order to avoid tuning instability, we evaluated the average of five runs (Hopkins and May, 2011). The results are summarized in Table 3. Our proposed methods achieved large gains in NTCIR10 task with the filtered method, but observed no gain in the KFTT with the filtered method. In NTCIR10 task with GDF, the gain in BLEU was smaller than that of KFTT. We calculate p-values and the difference between symmetric and c2c (the most effective proposed constraint) are lower than 0.05 in kftt with GDF and NTCIR10 with filtered method. There seems to be no clear tendency in the improved alignment qualities and the translation qualities, as shown in numerous previous studies (Ganchev et al., 2008).

Table 3: Results of translation evaluation

	KFTT		NTCIR10	
	GDF	Filtered	GDF	Filtered
symmetric	19.06	**19.28**	*28.3*	29.71
f2f	*19.15*	19.17	**28.36**	*29.74*
c2c	**19.26**	19.02	**28.36**	**29.92**
f2c	18.91	*19.20*	**28.36**	29.67

5 Conclusion

In this paper, we proposed new constraint functions under the posterior regularization framework. Our constraint functions introduce a fine-grained agreement constraint considering the frequency of words, a 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 of Setiawan et al. (2007). Experiments on word alignment tasks showed better alignment qualities measured by F-measure and AER on both the Hansard task and KFTT. We also observed large gain in BLEU, 0.2 on average, when compared with the previous posterior regularization method under NTCIR10 task.

As our future work, we will investigate more precise methods for deciding function words and content words for better alignment and translation qualities.

References

Arianna Bisazza and Marcello Federico. 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*, pages 439–448. 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.

Yin-Wen Chang, Alexander M. Rush, John DeNero, and Michael Collins. 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)*, pages 1481–1490, Baltimore, Maryland, June. Association for Computational Linguistics.

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. Association for Computational Linguistics.

John DeNero and Klaus Macherey. 2011. Model-based aligner combination using dual decomposition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 420–429, Portland, Oregon, USA, June. Association for Computational Linguistics.

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 Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 99:2001–2049.

Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K Tsou. 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*.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

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, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. 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.

E. Matusov, R. Zens, and H. Ney. 2004. Symmetric Word Alignments for Statistical Machine Translation. In *Proceedings of COLING 2004*, pages 219–225, Geneva, Switzerland, August 23–27.

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 the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Hendra Setiawan, Min-Yen Kan, and Haizhou Li. 2007. Ordering phrases with function words. In *Proceedings of the 45th annual meeting on association for computational linguistics*, pages 712–719. Association for Computational Linguistics.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*.

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.