Building a Better Bitext for Structurally Different Languages Through Self-training

Jungyeul Park* **Loïc Dugast†** **Jeen-Pyo Hong‡** **Chang-Uk Shin§** **Jeong-Won Cha§**
*Department of Linguistics. University of Arizona. Tucson, AZ 85721
†Academy of African Language and Science, University of South Africa, South Africa
‡NAVER Corporation, Republic of Korea
§Department of Computer Engineering, Changwon National University. Republic of Korea
http://air.changwon.ac.kr

Abstract

We propose a novel method to bootstrap the construction of parallel corpora for new pairs of structurally different languages. We do so by combining the use of a pivot language and self-training. A pivot language enables the use of existing translation models to bootstrap the alignment and a self-training procedure enables to achieve better alignment, both at the document and sentence level. We also propose several evaluation methods for the resulting alignment.

1 Introduction

A parallel corpus is a pair of texts written in different languages which are translation of each other. Since multilingual publication has become more widespread, there is an increasing amount of such parallel data available. Those are valuable resources for linguistic research and natural language processing applications, such as machine translation. It is also valuable when building cross-lingual information retrieval software. Finding the corresponding documents between two languages is a required step to build a parallel corpus, before more fine-grained alignments (paragraphs and sentences) can be calculated. In some scenarios, multilingual data with identical or considerably similar texts can be found with more than two languages involved. We ask whether a language can help as a pivot when aligning corpora and whether self-training may bring additional improvement of the alignment quality. We see further that both questions can be answered positively.

We propose a novel method to efficiently build better parallel corpora through the combination of pivot language and self-training. This method is especially targeted at aligning structurally different languages. We present a topic-based document alignment algorithm and a length and lexicon-based sentence alignment algorithm. Instead of directly aligning languages with widely different structures and even different writing systems, we make use of a pivot language and translate the other language into this pivot language before performing alignment. Translation can be done with a statistical translation model if previous existing parallel data exist. In our case, we perform a joint alignment and training of a translation model for the Korean-English language pair. We use English as a pivot language. Therefore, Korean sentences are translated into English before getting aligned. That is, we align English and English-translated Korean instead of directly aligning English and Korean. In the end, alignments are restored in the original languages to build a parallel corpus. We also employ a self-trained translation model in which the statistical translation model is reinforced by the newly aligned data.

The contribution of this work is mainly as follows: (1) We use a pivot language to align two languages with different writing systems. (2) We propose a self-training method to be able to produce better parallel corpora. (3) We describe the basic preprocessing scheme for Korean to be able to improve the statistical machine translation results. (4) We also propose several experiments for aligned parallel corpora by providing a standard

* Most work has been done when J. Park was at the University of Arizona and L. Dugast was at the University of South Africa. Current J. Park's affiliation is CONJECTO, Rennes, France and L. Dugast's affiliation is TextMaster, Paris, France.

Proceedings of the Workshop on Curation and Applications of Parallel and Comparable Corpora, pages 1–10,
Taipei, Taiwan, November 27–December 1, 2017.

evaluation data set for Korean. We hope that the present work will pave the way for further development of machine translation for Korean.

2 Case Study for Crawling Parallel Documents from the Web

When we try to build a good parallel corpus by crawling bilingual (or multilingual) documents from the Web, we may encounter unexpected difficulties. In this section, we show a case study to point out these difficulties in building a parallel corpus for Korean using bilingual documents crawled from the Web. We obtain the bilingual data from the KOREANA website, a quarterly journal published on-line.[1] It offers information on Korean culture, originally written in Korean, along with their translations into several languages. For our small experiments in this case study, we work on web pages written in Korean and their translations into English. We first align documents then sentences. We crawl and prepare 348 Korean and 381 English documents of the time-span (2005-2014). Sentences in (1-4) extracted from a document of the KOREANA site, show the example results of alignment by our proposed method (alignment through translation and self-training) as described in §3.

After aligning documents and sentences, results on Korean-English machine translation do not improve when using the newly produced aligned corpus. Actually, even though they present relatively *good* quality of document and sentence alignments, we notice that all English sentences do not exactly correspond to Korean sentences, but are rather loose translation of them or even involve substantial rewriting. Mismatches of the words in the aligned sentences are represented in gray. We also estimate their correctness of translation by a ratio which we simply calculate based on the number of correctly translated words into English and the number of correctly translated words from Korean as follows:

$$\text{Correctness of translation} = \frac{\text{\# of correctly translated words}}{\text{Total \# of words}} \quad (1)$$

where # are the number of words in Korean and English. Such mismatches in the aligned corpus will generate in bad quality of the translation model. We estimate that over half of English sentences are not exactly translated from Korean.

[1] http://www.koreana.or.kr

description	notation
KO corpus	C_k
EN translated KO corpus	$C^i_{k'}$
EN corpus	C_e
Bilingual KO-EN	BC^i_{KoEN}
KoEN MT system	$MT(\sum_i BC^i_{\text{KoEN}})$

Table 1: Notations for the Bilingual Setting

Therefore, even though we can align correctly such a corpus at the sentence level, we may not obtain good quality of the translation model. Actually, many sites which provide bilingual (or multilingual) language services, especially translated from Korean into other language, show similar characteristics. We consider that they are rather comparable corpora and it would be difficult to expect good quality of sentence-aligned data from these sites. Working on comparable corpora is beyond the scope of this paper.

3 Proposed Method

Notations for this self-training setting are described in Table 1.

3.1 Document alignment

For the document alignment task, we make the hypothesis that some topics are similar or even identical between the original and its translations. We can therefore make use of a topic model to find the similarity between two documents. Probabilistic topic models enable to discover the thematic structure of a large collection of documents. It provides a latent topic representation of the corpus. Latent Dirichlet Allocation (LDA) is one of the most used type of topic models (Blei et al., 2003). In LDA, a document may be viewed as a mixture of topics and represented as a vector. This enables the comparison of document topics in a vector space.

The cosine similarity measure is applied to two latent vectors of documents in different languages. Let $similarity(d_{L_1}, d_{L_2})$ the cosine similarity between two documents in two different languages L_1 and L_2. This cosine similarity is calculated as follows:

$$similarity(d_{L_1}, d_{L_2}) = \frac{V_{d_{L_1}} \cdot V_{d_{L_2}}}{\|V_{d_{L_1}}\| \|V_{d_{L_2}}\|} \quad (2)$$

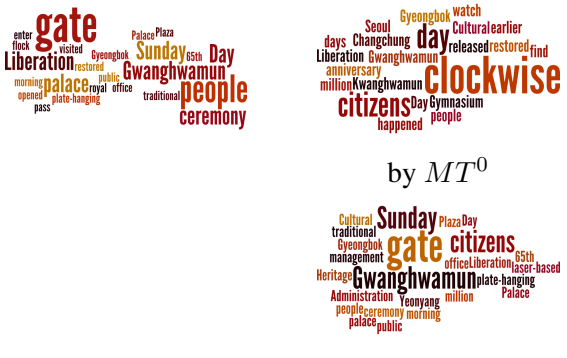
where two word vectors of $V_{d_{L_1}}$ and $V_{d_{L_2}}$ are from two documents in L_1 and L_2 languages. Instead of

(1) a. 그러나 경복궁에는 조선 창업의 뜻이 담겨 있으며, 500여 년 동안 조선을 상징하는 장소로 인식되었다.
b. Still, Gyeongbokgung does embody the spirit of the Joseon founders and for some 500 years has stood as an enduring symbol of the Joseon dynasty. (97.01%)

(2) a. 그러한 경복궁에 일본 식민지 통치를 위한 중추 기관인 조선총독부 신청사를 건설한 것은 지독히도 폭력적인 방법이었다.
b. Since Korea's liberation in 1945, there had been calls for the removal of the government general's building, which served as a painful reminder of Japan's colonial rule. (64.47%)

(3) a. 1990년대 조선총독부 건물을 헐어내고 경복궁을 복원하기 시작한 것은 사실 매우 논란의 여지가 있는 작업이다.
b. But upon the demolition of this building in the early 1990s, which enabled the Gyeongbokgung restoration project to get underway, even this was not free of its own controversy. (63.51%)

(4) a. 그러나 이러한 기억 투쟁이 식민지 시기를 떨쳐내고자 하는 사회적 요구에 의한 것이라는 점도 부인할 수는 없다.
b. In any case, no one can dispute the value of restoring Gyeongbokgung to its former glory and magnificence. (18.75%)



by MT^0

by MT^1

(a) English (b) Translated Korean

Figure 1: Examples of topic models in English and translated English from Korean

using all words in the document, we build these vectors from the topic models described above. Given a document in Korean and English, we translate them into English using the trained statistical translation models. We know that original Korean and English topic models do not directly share their elements in a vector space. However, translated Korean and English data by MT^0 show increasingly similar topic models and they become visibly related to each other. This situation improved further after self-training by MT^1 (See Figure 1). Measuring such similarity is hardly possible without using a pivot language or translated resources (Wu and Wang, 2007, 2009).[2]

3.2 Sentence alignment

Sentence alignment has been well-studied in the early 1990s (Brown et al., 1991; Chen, 1993; Gale and Church, 1993; Kay and Roscheisen, 1993). However, development of machine translation research demands increasing volumes of parallel data. This situation has led to the reinvestigation of sentence alignment such as in Moore (2002) and Varga et al. (2005) during the last decade. Actually, many sentence alignment methods were designed for related languages. The length-based alignment method in Gale and Church (1993) was originally intended for the related languages of English, French and German. The method of Kay and Roscheisen (1993) which uses a partial alignment of lexical items (cognates) to perform sentence alignment is also meant to be used for languages close to each other in the phylogenetic sense.

But directly aligning fairly dissimilar languages with different writing systems still remains a challenging task. For example, this could explain why the size of the Greek-English parallel corpus is one of the smallest corpora for the same timespan (1996-2011) in the Europarl Parallel Corpus, since the Greek language does not share the writing system of the other languages in the European Union. To explore the alignment of languages using different writing systems, Wu (1994) applies the method of Gale and Church (1993) to a parallel corpus between Cantonese and English from the Hong Kong Hansard using lexical cues, and Haruno and Yamazaki (1996) which is a variant of Kay and Roscheisen (1993) uses statistical and dictionary information for a parallel corpus between Japanese and English.

Accordingly, Moore (2002) and Varga et al. (2005) used partial translation models. Moore (2002) introduced a modified version of the well-known IBM Translation Model 1 using the highest probability 1-to-1 bids from the initial alignment. Varga et al. (2005) produced a crude translation

[2] See Zhang et al. (2017), as an exception.

of the source text using an existing bilingual dictionary. It seems natural that a translation model should be of precious help to align languages with different writing systems.

In this paper, we extend the length-based Gale and Church sentence alignment algorithm. The proposed algorithm is detailed in Hong (2013). Let $D(i, j)$ be the minimum distance. This is computed by minimizing operations as defined in Gale and Church (1993). We use the distance function d with six arguments s_1, t_1, s_2, t_2, s_3, t_3 instead of first four arguments. This is to extend to grouping up to three sentences, instead of two. Semantics of calculating $d(\cdot)$ is described in Figure 2. For example, $d(s_1, t_1; s_2, t_2; s_3, t_3)$ designates the cost of merging s_1, s_2, s_3 and matching with t_1, t_2, t_3. $\lambda_1 = 0.04, \lambda_2 = 0.21, \lambda_3 = 0.75$ are empirically estimated from the existing English-Korean parallel corpus, where $\sum_i \lambda_i = 1$.

3.3 Self-training method

We use a translation model learned from a previous alignment to produce an improved alignment at both document and sentence levels. This kind of practice is often called self-training (McClosky et al., 2006), self-taught learning (Raina et al., 2007), and lightly-supervised training (Schwenk, 2008). We assume that the initial, baseline translation models are trained with "out-domain" corpus, while the self-trained models are trained with "in-domain" corpus. Self-training therefore performs domain-adaptation that is beneficial to the quality of the final alignments.

At first, we translate Korean (C_k into English ($C^0_{k'}$) using the machine translation (MT) system trained with the pre-existing Korean-English bilingual corpus, as noted by $MT(BC^0_{\text{KoEN}})$. We then align documents and sentences to produce the parallel text for *translated* Korean and English. By restoring the original Korean sentences from translated Korean ($C^0_{k'}$) we build a new parallel corpus (BC^1_{KoEN}). From here, we can train a new MT system by adding the newly aligned bilingual corpus ($MT(BC^0_{\text{KoEN}} + BC^1_{\text{KoEN}})$) and re-translate Korean into English to build a self-trained BC^2_{KoEN}. This procedure can be summarized as follows:

1. Build a translation model using the existing parallel corpus $MT(BC^0_{\text{KoEN}})$.
2. Translate Korean C_k into English $C^0_{k'}$ using $MT(BC^0_{\text{KoEN}})$.
3. Align $C^0_{k'}$ and C_e.
4. Restore $C^0_{k'}$ to Korean and create a new parallel corpus BC^1_{KoEN}.
5. Build a new translation model by adding the newly aligned parallel corpus $MT(BC^0_{\text{KoEN}} + BC^1_{\text{KoEN}})$.
6. Repeat from (2) to (4) to create a self-trained parallel corpus BC^2_{KoEN}.

Through self-training, we can improve the translation quality for $C^i_{k'}$ and finally obtain better alignment results. Therefore, $C^i_{k'}$ (translation by $MT(\sum_i BC^i_{\text{KoEN}})$) and BC^{i+1}_{KoEN} are the corpora produced during self-training where $i = 0, 1$.

Figure 3 shows examples of English-Korean self-training. It shows their *intermediate* translation for original Korean sentences by the initial translation model and self-trained translation model. It is clear that the self-trained translation model is reinforced by the previously aligned corpus in which it provides more context-proper translation.

4 Experiments and Results

In this section, we detail our experiments and present our alignment results obtained through machine translation and self-training[3].

4.1 Data and systems

We experiment on a corpus extracted through web crawling. The corpus consists of news-wire articles from the *Dong-a Ilbo* website (literally 'East Asia Daily'). We obtained articles published during 2010 and 2011. It amounts to 3,249 documents for both Korean and English, containing 47,069 and 46,998 sentences respectively.

As far as non-linguistic preprocessing is concerned, we perform corpus cleaning using simple regular expressions after detecting text bodies. Since most contemporary HTML documents are created and edited by an HTML-specialized editor, we can easily detect the beginning and the end of text bodies in the document. Then, we can use the following regular expression to remove remaining HTML tags: `cat filename | sed "s/<[^>]*>//g"`. We empirically found that

[3] All obtained aligned data including source data (non-aligned original data) are made publicly available for further research.

$$D(i,j) = \min \begin{cases} D(i,j-1) + d(0,t_j;0,0;0,0) \\ D(i-1,j) + d(s_i,0;0,0;0,0) \\ D(i-1,j-1) + d(s_i,t_j;0,0;0,0) \\ D(i-1,j-2) + d(s_i,t_j;0,t_{j-1};0,0) \\ D(i-2,j-1) + d(s_i,t_j;s_{j-1},0;0,0) \\ D(i-2,j-2) + d(s_i,t_j;s_{j-1},t_{j-1};0,0) \\ D(i-1,j-3) + d(s_i,t_j;0,t_{j-1};0,t_{j-2}) \\ D(i-3,j-1) + d(s_i,t_j;s_{j-1},0;s_{j-2},0) \\ D(i-2,j-3) + d(s_i,t_j;s_{j-1},t_{j-1};0,t_{j-2}) \\ D(i-3,j-2) + d(s_i,t_j;s_{j-1},t_{j-1};s_{j-2},0) \\ D(i-3,j-3) + d(s_i,t_j;s_{j-1},t_{j-1};s_{j-2},t_{j-2}) \end{cases}$$

$$d(s_1,t_1;s_2,t_2;s_3,t_3) = \lambda_1 \log_2 Prob(\delta|match) + \lambda_2 + \lambda_3 \text{ cosine}(s_1+s_2+s_3, t_1+t_2+t_3)$$

Figure 2: Minimum distance

문화재청이 복원된 광화문의 현판 제막식을 갖고 광화문을 공개한 15일 광화문을 관람하려는 시민들의 발길이 하루 종일 이어졌다.
이날 오전 서울 광화문광장에서 열린 제65주년 광복절 경축식이 끝난 뒤 광화문을 지나 경복궁으로 들어가고 있는 시민들.
경복궁관리소는 이날 광화문을 찾은 시민이 10만여 명에 달한다고 밝혔다.

(a) Original Korean sentences

The Cultural restored Kwanghwamun's happened 제막식 Gwanghwamun, and released clockwise to watch the 15 days to citizens of 종일 a day.
Earlier in the day, Seoul's 광화문광장 65 anniversary of Liberation Day Changchung Gymnasium for after clockwise to Gyeongbok into and citizens.
The 경복궁관리소 clockwise to find the 10 million people.

(b) Translation from Korean into English by the initial MT model (MT^0)

The Cultural Heritage Administration laser-based traditional gate of a plate-hanging ceremony Sunday morning to the public 15 Gwanghwamun on to citizens of the Yeonyang.
At Gwanghwamun Plaza, the 65th Liberation Day after the gate into Gyeongbok Palace, and citizens.
The palace's management office under the the gate 10 million people 201,800 said.

(c) Translation after self-training (MT^1)

People flock to the restored gate of Gwanghwamun on Liberation Day Sunday.
The royal palace gate was opened to the public after a plate-hanging ceremony in the morning.
After a ceremony for the 65th Liberation Day at Gwanghwamun Plaza, people pass the traditional gate to enter Gyeongbok Palace.
The palace's office said more than 100,000 people visited the gate.

(d) Original English sentences

Figure 3: Examples of self-training

the proposed regular expression followed by manual detection of text bodies performs better than that the use of specific web page cleaning tools. This is especially true for web pages of *Dong-a Ilbo*, which require only one iteration of manual tagging, we can easily detect body parts which have the same structures for all documents. However, our method can be generalized by using such tools in future research.

After extracting text parts, sentence boundaries are detected using the ESPRESSO[4] POS tagger for Korean and SPLITTA described in Gillick (2009)[5] for English. We use these sentence segmented documents for document and sentence alignments. Then, we tokenize sentences using different methods depending on the language. As described before, we use the POS tagging system to tokenize Korean sentences and during the sentence segmentation task, tokenization is also performed. We use MOSES's tokenization script for English sentences. We also change the case of letters based on true case models for English.

For document alignment, we use LDA implemented in MALLET[6] to extract topic models. We convert the topics of each document into a single vector. We measure cosine similarity between two documents in different languages. Since we are working on English and English-translated Korean, we don't need polylingual topic models. For sentence alignment, we use a sentence alignment tool based on Hong (2013), which extends the algorithm of Gale and Church (1993). This sentence aligner enables the alignment of translated sentences and to restoration of original sentences based on sentence positions.

For Korean-English translation, we build the initial phrase-based statistical machine translation system using Korean parallel data that we previously collected from several bilingual Korean newswire sites. We do so with the Moses (Koehn et al., 2007) toolkit.[7] For alignment, we limit sentence length to 80 and use GIZA++ (Och and Ney, 2003). We use the SRILM (Stolcke, 2002) toolkit with Chen and Goodman's modified Kneser-Ney discounting for 5-grams for language model estimation. We also use `grow-diag-final-and` and `msd-bidirectional-fe` heuristics.[8] Finally, we use minimum error rate training (MERT) (Och, 2003) to tune the weights of the log-linear model.

[4] https://doi.org/10.5281/zenodo.884606

[5] https://code.google.com/p/splitta

[6] http://mallet.cs.umass.edu

[7] While we tested with a neural MT (NMT) system (Klein et al., 2017), the proposed method by SMT outperformed results from state-of-the-art NMT, most likely because of the small size of parallel data. We leave for future work the comparison of performance/results between statistical and neural systems with a bigger English-Korean bitext.

	Korean	MT^0	MT^1
precision	-	0.9701	0.9987
recall	-	0.9408	0.9981
F_1	-	0.9552	0.9984

Table 2: Results on document alignment

4.2 Results on document alignment

For the evaluation of document alignment, we use the name of documents as gold standard. Since the name of documents are identical for the Korean-English paired documents, for example 20101003K for Korean and 20101003E, we use this information as gold reference. Results on document alignment presented in this section are purely based on our proposed method that makes use of a topic model without referring to the name of documents. We evaluate our proposed methods using standard precision and recall as follows:

$$\begin{aligned} &\text{Precision} \\ &= \frac{\text{\# of correctly paired documents}}{\text{\# of produced alignment by threshold}} \\ &\text{Recall} \\ &= \frac{\text{\# of correctly paired documents}}{\text{\# of total paired documents}} \end{aligned} \tag{3}$$

We report F_1 score based on precision and recall ($\frac{2PR}{P+R}$). Table 2 shows results on document alignment. We denote MT^0 for $MT(BC^0_{\text{KoEN}})$ and MT^1 for $MT(BC^0_{\text{KoEN}} + BC^1_{\text{KoEN}})$ for convenience' sake. We introduce a threshold $\theta \geq 0.5$ of similarity for document alignment. Empirically we found that the recall drops if the threshold is set too high. For example, obtaining a precision of 1 comes with a drop in recall of 25% from $\theta \geq 0.7$ to ≥ 0.8. By using the proposed method, we obtain up to 99.84% F1 score.

4.3 Results on sentence alignment

To evaluate sentence alignment, we manually align sentences to build a gold standard. We se-

[8] http://www.statmt.org/moses/?n=Moses.Baseline for more details.

	Korean	MT^0	MT^1
sent	37,333	39,209	38,802
tok	1,193,514	1,193,509	1,193,507

Table 3: Size of sentence alignment: (sent) for the number of sentences and (tok) for tokens in the English-side corpus.

	Korean	MT^0	MT^1
P	0.4943	0.5547	0.5575
R	0.5385	0.5874	0.5927
F_1	0.5154	0.5705	0.5746

Table 4: Results on sentence alignment

lect documents over a period of two months (documents from March and April 2010). It contains over 1,500 sentences for each language from 122 documents. We evaluate our proposed methods using precision and recall as before:

$$P = \frac{\text{\# of correct bids}}{\text{\# of produced bids}}, R = \frac{\text{\# of correct bids}}{\text{\# of total bids}} \quad (4)$$

Table 3 shows the size and results on sentence alignment. We report overall precision, recall and F_1 scores. We provide results on sentence alignment without translation in which sentence alignment is based on sentence length only (Korean). MT^0 is for alignment by translation and MT^1 is for alignment by self-training. Table 5 present results for each bid by MT^1 and their occurrences in the evaluation data. Bids represent Korean:English. We found that many Korean sentences are not translated into English and the proposed sentence alignment method can correctly detect them. Some errors occur in 1:1 bids because the alignment method have a tendency to merge adjacent sentences, it can show better results in higher bids such as $n : m$ where $n, m > 1$.

Finally, we perform an extrinsic evaluation of alignment quality by evaluating a machine translation system. We train with the newly aligned corpus and evaluate the translation model using the JHE evaluation data (Junior High English evaluation data for Korean-English machine translation)[9] and the Korean-English News parallel corpus[10]. The direction of translation is Korean into English. Table 6 shows results using the translation quality metric BLEU (Papineni et al., 2002)[11].

5 Discussion on the Proposed Method

In this section, we first discuss the generalization of our proposed method, so that it does not get limited to the current bilingual setting. In the multilingual setting, we assume that we aim at aligning the source language and any other target language. We assume that there is a pivot language. Notations for this trilingual setting are described in Table 7. We use some analogy that we described for the bilingual setting in Table 1, such as C_k for the source language corpus (*e.g* Korean), C_e for the pivot language (English), and in addition C_f for the target language corpus (say, French).

Let k and f be Korean and French, respectively. English is a pivot language. We can use the result from the bilingual setting for the Korean to English translation to translate Korean into English. Then, we translate French into English using a MT system trained with a pre-existing French-English bilingual corpus. Finally, we align documents and sentences using English translated Korean-French documents to produce the parallel corpus by restoring the original Korean and French sentences. In the trilingual setting, we can also align French and English to improve the translation quality from French into English by providing a self-trained aligned corpus as we perform for Korean-English alignment. This procedure can be summarized as follows:

1. Create a self-trained parallel corpus BC^n_{KE} using the bilingual setting and build a translation model MT^n_{KE}.
2. Translate Korean C_k into English $C_{k'}$ using MT^n_{KE}.
3. Build a translation model using the existing parallel corpus $MT(BC^0_{\text{FE}})$.
4. Translate French C_f into English $C^0_{k'}$ using $MT(BC^0_{\text{FE}})$.
5. Align $C_{k'}$ and $C^0_{f'}$.
6. Restore $C_{k'}$ and and $C^0_{f'}$ to Korean and create a new parallel corpus BC^1_{KF}.

[9] https://doi.org/10.5281/zenodo.891295

[10] https://github.com/jungyeul/korean-parallel-corpora

[11] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a-20091001.tar.gz

	0:1	1:0	1:1	1:2	1:3	2:1	2:2	2:3	3:1	3:2	3:3
F_1	1.0	1.0	0.3552	0.7350	0.625	0.4761	0.8333	0.5	0.6667	1.0	1.0
occurrences	2	36	822	117	8	42	18	2	3	2	2

Table 5: Final results on sentence alignment for each bid for MT[1]

	Ko	MT^0	MT^1	
w/o (BC^0_{KoEn})	4.10	4.39	4.55	JHE
with (BC^0_{KoEn})	7.47	8.03	**8.33**	JHE
with (BC^0_{KoEn})	9.17	9.35	**9.38**	News

Table 6: Results on sentence alignment by BLEU scores. Ko is for results of the baseline system where the corpus is aligned with the pivot language. We also perform the translation with and without the initial bilingual corpus BC^0.

7. Align $C_{f'}$ and C_e.
8. Restore $C^0_{f'}$ to French and create a new parallel corpus BC^1_{FE}.
9. Build a new translation model by adding the newly aligned parallel corpus $MT(BC^0_{\text{FE}} + BC^1_{\text{FE}})$.
10. Repeat from (3) to (9) to create a self-trained parallel corpus BC^i_{KF}.

Through self-training, we can improve the translation quality for $C_{f'}$ by using the self-trained French-English parallel corpus BC_{FE}. Finally, we obtain better alignment results between Korean and French thanks to the better translation $C_{f'}$. Practically, it would be difficult to apply the proposed generalized method to real data because of the lack of proper multilingual data for Korean. We are aware that there are some multilingual data for Korean such as technical documents and movie/tv-show subtitles (Some of them are already available at OPUS).[12] According to our previous experience, these types of corpora are relatively easy to align because they may contain lexical cues (technical terms) or time stamps (subtitles).

6 Conclusion and Future Perspectives

We explored the possibility of using a pivot language for the purpose of aligning two dissimilar languages. Results show that alignment as evaluated directly by document and sentence alignments or indirectly by translation quality (BLEU), is improved as compared with directly aligning those two languages. Applying the generalized method for other language pairs such as Greek-English in the Europarl parallel corpus, in which multilingual parallel data are available and Greek does not share the same writing system with other European languages, can be considered as future work. In addition to using the pivot language, we also built a better parallel corpus using self-trained translation models. For immediate future work, we continue to identify suitable bilingual/multilingual web sites to collect more parallel data for Korean.

[12] http://opus.lingfil.uu.se

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and effort to improve the manuscript. J. Park and L. Dugast would like to thank Kyung Min Shin for the KOREANA bilingual data. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A1B03033534) for C.-U. Shin and J.-W. Cha.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3(4-5):993–1022.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Berkeley, California, USA, pages 169–176. https://doi.org/10.3115/981344.981366.

Stanley F. Chen. 1993. Aligning Sentences in Bilingual Corpora using Lexical Information. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Columbus, Ohio, USA, pages 9–16. https://doi.org/10.3115/981574.981576.

Description	Notation
Source language corpus	C_k
Pivot language translated source language corpus	$C^i_{k'}$, where $0 \leq i < n$
Target language corpus	C_f
Pivot language translated target language corpus	$C^i_{f'}$, where $0 \leq i < n$
Pivot language corpus	C_e
Bilingual Source-Pivot corpus	BC^i_{KE}, where $0 \leq i \leq n$
Bilingual Target-Pivot corpus	BC^i_{FE}, where $0 \leq i \leq n$
Bilingual Source-Target corpus	BC^i_{KF}, where $0 < i \leq n$
Source-Pivot MT system	$MT(\sum_i BC^i_{\text{KE}})$
Target-Pivot MT system	$MT(\sum_i BC^i_{\text{FE}})$

Table 7: Notations for the Multilingual Setting

William A. Gale and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* 19(1):75–102.

Dan Gillick. 2009. Sentence Boundary Detection and the Problem with the U.S. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Association for Computational Linguistics, Boulder, Colorado, pages 241–244. http://www.aclweb.org/anthology/N/N09/N09-2061.

Masahiko Haruno and Takefumi Yamazaki. 1996. High-Performance Bilingual Text Alignment Using Statistical and Dictionary Information. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Santa Cruz, California, USA, pages 131–138. https://doi.org/10.3115/981863.981881.

Jeen-Pyo Hong. 2013. *Multilingual sentence alignment using translation models*. Ph.D. thesis, Changwon National University.

Martin Kay and Martin Roscheisen. 1993. Text-Translation Alignment. *Computational Linguistics* 19(1):121–142.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, Vancouver, Canada, pages 67–72. http://aclweb.org/anthology/P17-4012.

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*. Association for Computational Linguistics, Prague, Czech Republic, pages 177–180. http://www.aclweb.org/anthology/P07-2045.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and Self-Training for Parser Adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sydney, Australia, pages 337–344. https://doi.org/10.3115/1220175.1220218.

Robert C. Moore. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In Stephen D. Richardson, editor, *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*. Springer-Verlag, Tiburon, CA, USA, pages 135–244.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sapporo, Japan, pages 160–167. https://doi.org/10.3115/1075096.1075117.

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*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. 2007. Self-taught Learning: Transfer Learning from Unlabeled Data. In *Proceedings of the 24th International Conference on Machine Learning*. ACM, New York, NY, USA, ICML '07, pages 759–766. https://doi.org/10.1145/1273496.1273592.

Holger Schwenk. 2008. Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. In *Proceedings of the International Workshop on Spoken Language Translation*. Hawaii, USA, pages 182–189.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002*. Denver, Colorado, pages 901–904.

Dániel Varga, Lázló Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP (Recent Advances in Natural Language Processing)*. Borovets, Bulgaria, pages 590–596.

Dekai Wu. 1994. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Las Cruces, New Mexico, USA, pages 80–87. https://doi.org/10.3115/981732.981744.

Hua Wu and Haifeng Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 856–863. http://www.aclweb.org/anthology/P07-1108.

Hua Wu and Haifeng Wang. 2009. Revisiting Pivot Language Approach for Machine Translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, Suntec, Singapore, pages 154–162. http://www.aclweb.org/anthology/P/P09/P09-1018.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial Training for Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 1959–1970. http://aclweb.org/anthology/P17-1179.