Method for Building Sentence-Aligned Corpus from Wikipedia

Keiji Yasuda [†,‡] **and Eiichiro Sumita** [†,‡]

[†] ATR Spoken Language Translation Research Laboratories [‡] National Institution of Information and Communications Technology

Abstract

We propose the framework of a Machine Translation (MT) bootstrapping method by using multilingual Wikipedia articles. This novel method can simultaneously generate a statistical machine translation (SMT) and a sentence-aligned corpus. In this study, we perform two types of experiments. The aim of the first type of experiments is to verify the sentence alignment performance by comparing the proposed method with a conventional sentence alignment approach. For the first type of experiments, we use JENAAD, which is a sentence-aligned corpus built by the conventional sentence alignment method. The second type of experiments uses actual English and Japanese Wikipedia articles for sentence alignment. The result of the first type of experiments shows that the performance of the proposed method is comparable to that of the conventional sentence alignment method. Additionally, the second type of experiments shows that we can obtain the English translation of 10% of Japanese sentences while maintaining high alignment quality (rank-A ratio of over 0.8).

Introduction

Wikipedia has articles in more than 200 languages, and it is one of the most varied language resources in the world. The current version of Wikipedia expresses multilingual relationships by only interlanguage links. Although Wikipedia has been used as a bilingual language resource in some natural language processing (NLP) researches (Erdmann et al. 2008), an interlanguage link is insufficient information for conducting some NLP researches such as corpus-based machine translation. Therefore, sentence-aligned parallel corpuses are required for these researches.

In this study, we propose a novel MT bootstrapping framework that can simultaneously generate a statistical machine translation (SMT) and a sentence-aligned corpus. SMT assist general users of Wikipedia by translating Wikipedia articles automatically. The sentence-aligned corpus assists NLP researchers by expanding the application of Wikipedia as a multilingual language resource.

Proposed Method

Figure 1 illustrates the framework of MT bootstrapping in the case of using English and Japanese Wikipedia. As shown in this figure, the MT bootstrapping method involves the following steps:

Step 1: Translate Wikipedia articles using an MT system[1].

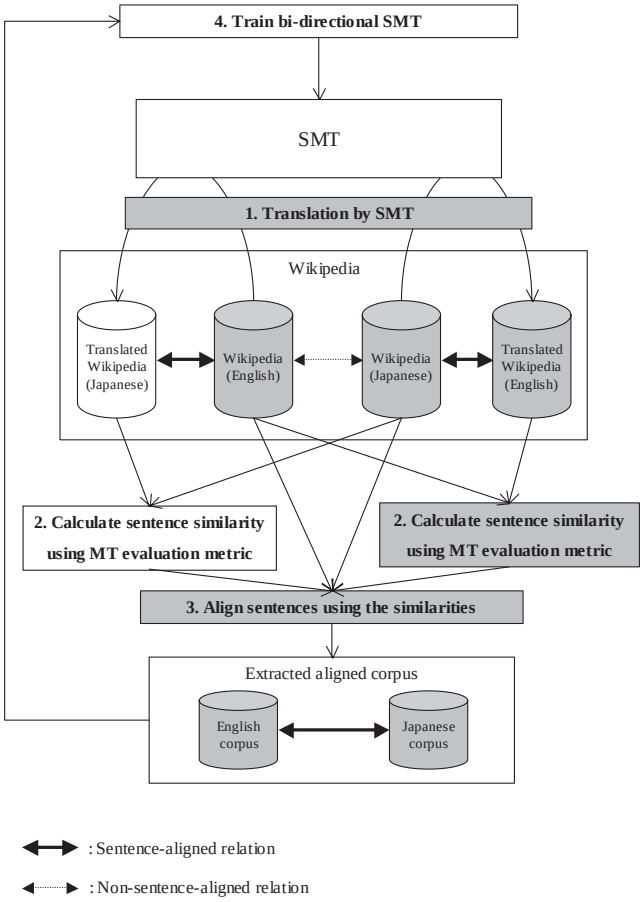

Fig. 1 MT bootstrapping framework

Step 2: Calculate the sentence-level MT evaluation score between Japanese sentences in the original Japanese Wikipedia and the Japanese sentences obtained by translating English Wikipedia. Similarly, calculate the sentence-level MT evaluation score between the target English sentences in the original English Wikipedia and the English sentences obtained by translating Japanese Wikipedia.

Step 3: Align sentence pairs from the original Japanese Wikipedia and English Wikipedia using either or both of the scores calculated in step 2.

Step 4: Train the SMT using the sentence-aligned corpus.

There are two main problems with this method. One is

[1] For the first loop, com mercial MT systems can be used instead of SMT.

the computational cost of implementing the above steps. Step 3 in particular requires excessive calculations because we have to calculate a sentence-level score for $2 \times n_{source} \times n_{target}$ pairs, where n_{source} is the number of sentences in the source language comparable corpus and n_{target} is the number of sentences in the target language comparable corpus.

One method to reduce the calculation cost is to use an interlanguage link to prune candidate articles. Then, a sentence-level bilingual evaluation understudy (BLEU) score can be calculated only for the candidate articles. This method effectively reduces the computational cost of step 3. However, the iteration of steps 1 to 4 still induces a large computational cost. To deal with these problems, we have raised funds and we use a supercomputer (HPC2500, 2002).

The other problem concerns the alignment confidence measure that is based on the MT evaluation measure. Most proposed MT evaluation measures are used to evaluate translation quality, and no research uses these measures for evaluating sentence alignment confidence. If our proposed framework functions effectively, it will be a promising application because most MT evaluation techniques do not require any language resources such as a bilingual thesaurus or lexicon, which are used in most conventional sentence alignment methods. In the experiments described in the next section, we test the effectiveness of the MT evaluation technique.

Experiments

We perform two types of experiments. The aim of the first type of experiments is to verify the sentence alignment performance by comparing the proposed method with a conventional sentence alignment approach. A JENAAD corpus, which is a sentence-aligned corpus built by the conventional sentence alignment method, is used for the experiments. The second type of experiments uses actual English Wikipedia and Japanese Wikipedia articles for sentence alignment.

Experiments using JENAAD corpus

Experimental settings. To verify the effectiveness of sentence alignment based on the MT evaluation measure, we perform the experiments depicted in Fig. 2 by using a preliminary sentence-aligned corpus. As shown in this figure, the experiments involve the following steps.

1. Train the SMT using the preliminary sentence-aligned corpus.

2. Translate the corpus using the trained translation system.

3. Calculate the sentence-level BLEU score (Papineni et al. 2002) between the preliminarily aligned sentence pairs.

4. Filter out unreliable aligned pairs using the BLEU score.

5. Retrain the STM using the filtered sentence pairs.

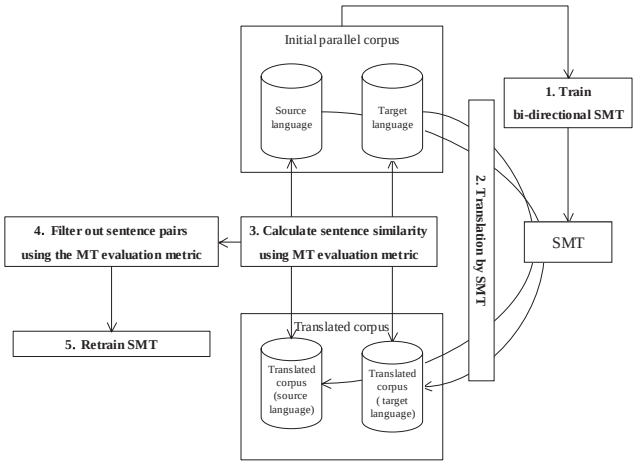

Fig. 2 Flow of JENAAD corpus experiments

To evaluate the effectiveness of the MT evaluation measure in sentence alignment, we compare the system performance before and after filtering. If the measure is useful for sentence alignment, we can reduce the size of the training set for the SMT without degrading the performance, by filtering out unreliable sentence pairs.

We used the JENAAD corpus, which is a Japanese-English newspaper corpus aligned by a conventional sentence alignment method (Utiyama & Isahara 2003). We use 150,000 sentence pairs from the corpus. For the SMT training, we use a Pharaoh training toolkit and an SRI language model toolkit.

Experimental results. Table 1 lists the results of the JENAAD corpus experiments. The BLEU score of the 500-sentence-pair test set is calculated in a manner similar to that in previous experiments. Here, the baseline system is trained on the sentence pairs that are filtered using the measure proposed by Utiyama et al. (Utiyama & Isahara 2003). As indicated in the table, the performance of the proposed method is comparable to that of the conventional sentence alignment measure.

Experiments using Wikipedia

Experimental settings. In the experiments using Wikipedia, we execute the first loop of the components shaded in gray[2], shown in Fig. 1, in order to verify the relationship between the yield ratio and the alignment quality. The other experimental conditions are as follows:

-The Wikipedia version is Sept. 2007.
-The number of sentences from Japanese Wikipedia is 1,500.
-The sentences of Japanese Wikipedia are preprocessed by a Chasen morphological analyzer.
-The number of sentences from English Wikipedia is 50,000,000.

[2] In these experiments, we used a commercial MT system instead of SMT.

Table 1 Results of JENAAD corpus

Method	# of the training sentence pairs	BLEU score
Proposed method	50000	0.1019
Proposed method	100000	0.109
Baseline	50000	0.1069
Baseline	100000	0.1057
Baseline	150000	0.1092

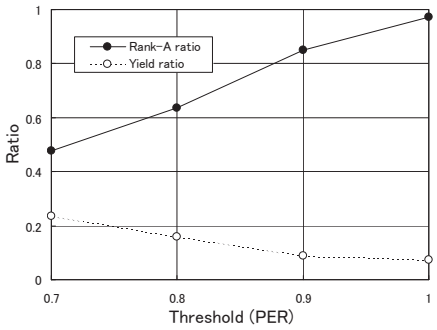

Fig. 3 Results of Wikipedia experiments

To align the Japanese and English sentences (step 3 in Fig. 1), first, we calculate the sentence similarity between Japanese sentences and all 50,000,000 Japanese translated sentences. We consider that original English sentence of a Japanese translated sentence, which gives the highest similarity value. For English sentences extracted from English Wikipedia articles as the translations of 1,500 Japanese sentences, we carry out a two-grade subjective evaluation using the following definition.

Rank A: more than 60% overlap.

Rank B: less than 60% overlap.

Experimental results. Figure 3 plots the results obtained by the Wikipedia experiments. In this figure, the vertical axis denotes the ratio and the horizontal axis shows the cutoff threshold for a translation pair. The blank circles in the figure indicate the yield ratio, which is the ratio of the obtained parallel sentences to all Japanese sentences (1,500 sentences). The filled circle indicates the rank-A ratio, which is the ratio of rank-A sentence pairs to all obtained sentence pairs. The results shown in the figure reveal that we can obtain the English translation of 10% of Japanese sentences while maintaining high alignment quality (rank-A ratio of over 0.8).

Related Works

Some previous researches have attempted to extract sentence pairs from comparable corpuses. Some of these researches require manually built language resources such as a bilingual thesaurus or lexicon to enhance the alignment performance (Utiyama & Isahara 2003; Ma 2006). However, our method requires only an initial sentence-aligned corpus.

There are two researches that propose concepts similar to our proposed concept (Fung & Cheung 2004; Munteanu & Marcu 2006). A common feature of these researches (Fung & Cheung 2004) is that they apply the bootstrapping framework for sentence alignment. Our proposed method uses a bilingual lexicon for sentence alignment and automatically updates the lexicon using the aligned sentences by the method by Fung and Cheung and that by Munteanu and Marcu These methods (Fung & Cheung 2004; Munteanu & Marcu 2006) use some elemental technology of STM to build a bilingual lexicon automatically; however, the entire STM technology has not been used. This is one of the differences between the proposed method and the conventional method.

Conclusions

We have proposed an MT bootstrapping method that simultaneously generates an STM and a sentence-aligned parallel corpus. This method iterates the following steps: (1) translation of a comparable corpus using the SMT, (2) sentence alignment of the comparable corpus using the MT evaluation measure, and (3) SMT training.

To test the effectiveness of the proposed method, first, we conducted preliminary experiments using a newspaper corpus. According to the experimental results, we thought that the sentence alignment based on the MT evaluation measure was effective and performed comparably to the conventional sentence alignment method.

Second, we performed sentence alignment experiments using English and Japanese Wikipedia articles.

The results of these experiments show that we can obtain the English translation of 10% of Japanese sentences while maintaining high alignment quality (rank-A ratio of over 0.8).

Future Works

Currently, we are performing actual MT bootstrapping experiments shown in Fig. 1 by using an HPC2500 supercomputer (HPC2500, 2002), which has 11 nodes with 128 CPUs in each node.

References

Erdmann, M., Nakayama, K., Hara, T., and Nishio, S. 2008. An Approach for Extracting Bilingual Terminology from Wikipedia, *Proc. of DASFAA-2008*.

HPC2500. 2002. http://pr.fujitsu.com/en/news/2002/08/22.html

Papineni, K., Roukos, S., Ward, T., and Zhu, W. 2000. Bleu: A Method for Automatic Evaluation of Machine Translation, *Proc. of ACL-2002*, pp. 311–318.

Utiyama, M., and Isahara, H. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. *Proc. of ACL-2003*, pp. 72–79.

Ma, X. 2006. Champollion: A Robust Parallel Text Sentence Aligner. *Proc. of LREC-2006*. pp.489-492.

Fung, P., and Cheung, C. 2004. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. *Proc. of EMNLP-2004*. pp.57-63.

Munteanu, D., and Marcu, D. 2006. Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora. *Proc. of ACL-2006*, pp. 81–88.