Inflating Training Data for Statistical Machine Translation using Unaligned Monolingual Data

Wei YANG, Zhongwen ZHAO, Yves LEPAGE
Graduate School of Information, Production and Systems, Waseda University
{kevinyoogi@akane; zzw890827@fuji}.waseda.jp; yves.lepage@waseda.jp

Abstract

In data-driven machine translation, parallel corpora are an extremely important resource. For language pairs that involve English, there exist many freely available bilingual or multilingual parallel corpora, especially for European languages. To improve the translation quality for less-resourced language pairs, such as Chinese–Japanese, larger and larger aligned training data are needed. The constitution of large bilingual corpora is not easy for less documented language pairs. In this paper, we show how to construct a Chinese–Japanese quasi-parallel corpus automatically by using analogical associations based on a small amount of parallel sentences and a reasonable amount of monolingual data. We perform SMT experiments in Chinese–Japanese and compare a baseline system and a system build by adding the quasi-parallel corpus. On the same test set, the translation quality significantly improved over the baseline system.

1 Introduction

Sentence-level aligned parallel corpora i.e., sets of parallel sentences, are an important resource as training data in statistical machine translation (SMT), because the necessary translation knowledge is acquired from these sentential parallel corpora, and because the translation relations between words or phrases between the source language and the target language are extracted from parallel sentences. There exist numerous freely available sentence-level bilingual or multilingual parallel corpora for language pairs that involve English, such as the Europarl parallel corpus [4]. There also exist some studies on parallel sentence extraction that include Chinese or Japanese as one of the languages involved. Examples are: parallel sentence mining from Chinese–English bilingual web pages [8]; extraction of Japanese-English parallel sentences from a noisy parallel corpus [7].

Our proposed method follows a slightly different path. We propose to automatically construct a bilingual corpus of *quasi-parallel* sentences based on a small amount of parallel corpus and a reasonable amount of unaligned monolingual data. A quasi-parallel corpus will contain aligned sentence pairs that are translations to each other to a certain extent. The motivation for our proposed method is that, while direct construction of large bilingual corpora is difficult, monolingual data is relatively easy to access in large amounts. The languages on which we will illustrate our proposed method in this paper are Chinese and Japanese.

The procedures in our experiments are as follows:

- Cluster and group monolingual data collected from the Web using *proportional analogy* independently in both Chinese and Japanese. Such clusters can be considered as *rewriting models* for new sentence generation. We also compute the similarity between these clusters across the two languages.
- Generate new sentences using these rewriting models starting from *seed sentences* extracted from the monolingual part of some parallel corpus, and filter out dubious sentences.
- Construct the quasi-parallel corpus by assessing the strength of translation relations between the filtered newly generated sentences based on the similarity between the clusters they were generated from and the translation relations between the seed sentences.
- Perform SMT experiments: compare the baseline system and a system build by adding the quasi-parallel corpus as the additional training data.

2 Data Preparation

2.1 Chinese–Japanese Parallel Sentences

The Chinese and Japanese linguistic resources we use in this paper are the ASPEC-JC (Asian Scientific Paper Excerpt Corpus-Japanese–Chinese) corpus[1]. This corpus is designed for Machine Translation and is split as follows:

- Training Data: 672,315 sentences;
- Development Data: 2,090 sentences;
- Test Data: 2,107 sentences.

For the monolingual part of the parallel sentences used in new sentence generation, we extracted short sentences with less than 30 characters in length from the Training

[1] http://orchid.kuee.kyoto-u.ac.jp/ASPEC/

Data, and obtained 103,629 sentences to be used as seed sentences for new sentence generation in both languages.

2.2 Chinese and Japanese Monolingual Sentences

For construction of the analogical clusters, we use Chinese and Japanese monolingual data. We collected Chinese and Japanese monolingual short sentences with less than 30 characters in size from the Web using an in-house Web-crawler, mainly from the following websites: "Yahoo China", "Yahoo China News", "douban" for Chinese and "Yahoo! JAPAN", "Mainichi Japan" for Japanese. We use 70,000 Chinese and Japanese monolingual data (after some filtering) in clustering experiments. Although the number of sentences in both language is the same, the sentences are basically unrelated.

3 Constructing Analogical Clusters

3.1 Sentential Analogies

We cluster and gather pairs of sentences in Chinese and Japanese independently using *proportional analogy*. Proportional analogies establish a structural relationship $A : B :: C : D$ between four objects, A, B, C and D: 'A is to B as C is to D'. An efficient algorithm for the resolution of analogical equations between strings of characters has been proposed in [5].

The algorithm relies on counting numbers of occurrences of characters and computing edit distances (with only insertion and deletion as edit operations) between strings of characters. For a detailed presentation of the formula and the algorithm, we refer the reader to [5]. We call *sentential analogy* a proportional analogy between sentences.

ご確認ください	:	ご了承ください	::	確認しました	:	了承しました
Confirm it, please.	:	*Understand it, please.*	::	*I have confirmed.*	:	*I have understood.*

3.2 Analogical Clusters

When several sentential analogies involve the same pairs of sentences, they form a series of analogous sentences, and they can be written on a sequence of lines where each line contains one sentence pair and where any two pairs of sentences form a sentential analogy. We call such a sequence of lines an *analogical cluster*. The following analogical cluster in Japanese shows three possible sentential analogies. Analogical clusters can be considered as rewriting models to generate new sentences.

Analogical clusters are constructed from the monolingual data separately in each language. Table 1 summarizes some statistics on the clusters produced.

ご 確認 お願いします	:	ご 了承 お願いします
'Please confirm it.'		*'Please understand it.'*
ご 確認 ください	:	ご 了承 ください
'Confirm it, please.'		*'Understand it, please.'*
確認 しました	:	了承 しました
'I have confirmed.'		*'I have understood it.'*

	Chinese	Japanese
# of different sentences	70,000	70,000
# of clusters	23,182	21,975

Table 1: Statistics on the Chinese and Japanese clusters constructed independently in each language.

3.3 Determining corresponding clusters by computing similarity

For determining corresponding clusters, we consider the changes between the left (S_{left}) and the right (S_{right}) sides in one cluster as two sets. We perform word segmentation[2] on these changes in sets to obtain minimal sets of changes made up with words or characters.

Then, we compute the similarity between the left sets (S_{left}) and the right sets (S_{right}) of Chinese and Japanese clusters. To this end, we make use of the EDR dictionary[3] and word-to-word alignments (based on ASPEC-JC data using Anymalign[4]), we keep 72,610 word-to-word correspondences obtained with Anymalign in 1 hour after filtering on both translation probabilities with a threshold of 0.3 (approximate correctness judged by hand: 95%). We also use a traditional-simplified Chinese variant table[5] and Kanji-Hanzi Conversion Table[6] to translate all Japanese words into Chinese, or convert Japanese characters into simplified Chinese characters. We calculate the similarity between two Chinese and Japanese word sets according to a classical Dice formula:

$$Sim = \frac{2 \times |S_{zh} \cap S_{ja}|}{|S_{zh}| + |S_{ja}|} \qquad (1)$$

S_{zh} and S_{ja} denote the minimal sets of changes across the clusters (both on the left or right) in both languages (after translation and conversion). To compute the similarity between two Chinese and Japanese clusters we take the arithmetic mean on both sides, as given in formula (2):

$$Sim_{C_{zh}-C_{ja}} = \frac{1}{2}(Sim_{left} + Sim_{right}) \qquad (2)$$

About 15,710 cluster pairs were judged valid corresponding clusters ($Sim_{C_{\mathrm{zh}}-C_{\mathrm{ja}}} \geq 0.300$) by the above

[2] Segmentation toolkits: Mecab: `http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html` for Japanese and Urheen, a Chinese lexical analysis toolkit (National Laboratory of Pattern Recognition, China) for Chinese.

[3] `http://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html`

[4] `http://anymalign.limsi.fr`

[5] `http://www.unicode.org/Public/UNIDATA/`

[6] `http://www.kishugiken.co.jp/cn/code10d.html`

steps. It is important to stress that 1/ the number of lines in corresponding clusters may be different, and 2/ the sentences in corresponding clusters are not translations in general, only the changes are in correspondence.

4 Generating New Sentences Using Analogical Associations

4.1 Generation of New Sentences

Analogy is not only a structural relationship. It is also a process [3] by which, “given two related forms and only one form, the fourth missing form is coined” [1]. If the objects *A*, *B*, *C* are given, we may obtain another object *D* according to the analogical equation $A : B :: C : D$. It can be illustrated with sentences:

$$\text{ご確認お願いします} : \text{ご了承お願いします} :: \text{あらかじめご確認ください} : x$$

$$\Rightarrow \ x = \text{あらかじめご了承ください}$$

Here, the solution of the analogical equation is $D =$ “あらかじめご了承ください” (Please understand it previously.). If we regard each sentence pair in a cluster as a pair $A : B$ (left to right or right to left), and any short sentence not belonging to the cluster as *C* (a *seed sentence*), a new sentence *D* can be forged.

4.2 Experiments on New Sentence Generation and Filtering

For the generation of new sentences, we made use of the clusters obtained in Section 3.2. The seed sentences are the monolingual part of Chinese and Japanese short sentences from the 103,629 ASPEC-JC parallel data. In this experiment, we generated new sentences with each pair of sentences in clusters for Chinese and Japanese respectively. We obtained about 105 million newly generated sentences for Chinese and about 80 million for Japanese. The grammatical quality, as assessed manually, is only of 29% and 40% of the sentences.

To filter out invalid and grammatically incorrect sentences and keep only well-formed sentences, we eliminate any sentence that contains an N-sequence of a given length unseen in the reference corpus. Similar techniques have been previously used [2]. In our experiment, we introduced begin/end markers to make sure that the beginning and the end of a sentence are also correct. The best quality was obtained for the values N=6 for Chinese and N=7 for Japanese on our reference corpora (about 1,700,000 sentences for both Chinese and Japanese). Quality assessment was performed on a sample of 1,000 sentences randomly and manually check by native speakers. The grammatical quality was at least 96%. For new valid sentences, we remember their corresponding seed sentence and the cluster they were generated from.

4.3 Deducing and Acquiring Quasi-parallel Sentences

We deduce translation relations based on the short parallel sentences for new sentence generation, and the correspondence between clusters in Chinese and Japanese. If the seeds of two new generated sentences in Chinese and Japanese are aligned in the short parallel sentences, and if the clusters which they were generated from are corresponding, we suppose that these two Chinese and Japanese newly generated sentences are translations of one another to a certain extent. We obtained 35,817 unique Chinese–Japanese aligned sentences. Among these aligned sentences, about 74% were found to be exact translations by manual check on a sampling of 1,000 pairs of sentences. This figure of 74% justify our use of the word *quasi-parallel* to characterize this corpus of aligned sentences.

5 SMT Experiments

5.1 Baseline Training Data = ASPEC-JC, seeds = short sentences in ASPEC-JC

To assess the contribution of the generated quasi-parallel corpus, we propose to compare two SMT systems. The first one is trained using the ASPEC-JC parallel corpus (baseline). The second one is trained on the ASPEC-JC data plus the additional quasi-parallel corpus. The segmentation tools used are Urheen for Chinese and Mecab for Japanese.

	Baseline	Chinese	Japanese
train	sentences	672,315	672,315
	words	18,847,514	23,480,703
	mean ± std.dev.	28.12 ± 15.20	35.05 ± 18.88
	+ Quasi-parallel	Chinese	Japanese
train	sentences	**708,132**	**708,132**
	words	19,175,105	23,867,354
	mean ± std.dev.	27.61 ± 15.33	34.47 ± 19.07
	Both experiments	Chinese	Japanese
tune	sentences	2,090	2,090
	words	60,458	73,177
	mean ± std.dev.	28.93 ± 15.86	35.01 ± 18.87
test	sentences	2,107	2,107
	words	59,594	72,027
	mean ± std.dev.	28.28 ± 14.55	34.18 ± 17.43

Table 2: Statistics when baseline training data = all sentences in ASPEC-JC.

We perform all experiments using the standard GIZA++/MOSES 1.0 pipeline [6]. As Table 3 shows, significant improvement over the baseline is obtained by adding the quasi-parallel generated data.

		BLEU
zh-ja	baseline	29.10
	+ additional training data	**32.03**
ja-zh	baseline	22.98
	+ additional training data	**24.87**

Table 3: Evaluation results when baseline training data = all sentences in ASPEC-JC. Boldfaced numbers stand for statistically significant improvements.

5.2 Baseline Training Data = seeds = short sentences in ASPEC-JC

In a second experiment, the training data is the same as the data used as seeds in sentence generation, i.e., we eliminated from the training data sentences with a length greater than 30 characters. Similarly, we inspect the impact of adding the quasi-parallel corpus.

	Baseline	Chinese	Japanese
train	sentences	103,629	103,629
	words	1,147,636	1,407,873
	mean ± std.dev.	11.19 ± 3.58	13.73 ± 4.12
	+ Quasi-parallel	Chinese	Japanese
train	sentences	**139,446**	**139,446**
	words	1,475,227	1,794,524
	mean ± std.dev.	10.82 ± 3.74	13.13 ± 4.35
	Both experiments	Chinese	Japanese
tune	sentences	2,090	2,090
	words	60,458	73,177
	mean ± std.dev.	28.93 ± 15.86	35.01 ± 18.87
test	sentences	2,107	2,107
	words	59,594	72,027
	mean ± std.dev.	28.28 ± 14.55	34.18 ± 17.43

Table 4: Statistics when baseline training data = short sentences in ASPEC-JC.

The experimental results are shown in Table 5.

		BLEU
zh-ja	baseline	18.49
	+ additional training data	**24.30**
ja-zh	baseline	15.89
	+ additional training data	**19.88**

Table 5: Evaluation results when baseline training data = short sentences in ASPEC-JC. Boldfaced numbers stand for statistically significant improvements.

6 Conclusions

In this paper, we presented a different way to inflate the training corpus used to improve the translation quality for the SMT system: we automatically generate a quasi-parallel corpus. The experimental data we use are ASPEC-JC corpus and certain amount of monolingual data collected from the Web. We use analogical associations to generate new sentences, and filter them to ensure grammatical correctness. The grammatical quality of the valid new sentences is at least 96%. We then assess translation relations between valid newly generated sentences across both languages, relying on the similarity between the clusters across languages. We automatically obtained a quasi-parallel corpus of 35,817 Chinese–Japanese sentences, 74% of which were found to be exact translations.

In SMT experiments, we inflated the training data with our quasi-parallel data in a rewarding way. On the same test set, the translation scores significantly improved over the baseline systems. It should be stressed that the data that allowed us to get such improvement are not so large in quantity and not so good in quality, but we were able to control both quantity and quality so as to consistently improve translation quality.

References

[1] Ferdinand de Saussure. *Cours de linguistique générale*. Payot, Lausanne et Paris, [1ère éd. 1916] edition, 1995.

[2] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Human Language Technology Conference (HLT2002)*, pages 128–132, San Diego, CA, USA, 2002. Morgan Kaufmann.

[3] Esa Itkonen. *Analogy as Structure and Process: Approaches in linguistics, cognitive psychology and philosophy of science*, volume 14. John Benjamins Publishing Company, Amsterdam / Philadelphia, 2005.

[4] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand, 2005.

[5] Yves Lepage and Etienne Denoual. Purest ever example-based machine translation: detailed presentation and assessment. *Machine Translation*, 19:251–282, 2005.

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

[7] Masao Utiyama and Hitoshi Isahara. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79, 2003.

[8] Zhenxiang Yan, Yanhui Feng, Yu Hong, and Jianmin Yao. Parallel sentences mining from the web. *Journal of Computational Information Systems*, 5(6): 1633–1641, 2009.