Sentence Generation by Analogy: Towards the Construction of A Quasi-parallel Corpus for Chinese-Japanese

Hao Wang, Wei Yang, Yves Lepage
Graduate School of Information, Production and Systems, Waseda University
{oko_ips@ruri; keinyoogi@akane; yves.lepage}.waseda.jp

Abstract

Parallel corpora are indispensable resources in data-driven approaches to machine translation: statistical and example-based. The major problem inherent in developing a Chinese-Japanese machine translation system is the lack of bilingual parallel corpus. We have implemented several based on proportional analogy techniques to produce new quasi-parallel sentences using monolingual data, collected from Web. By investigating the performance of new parallel sentences generation, this paper describes the experiments and efforts we made towards building a Chinese-Japanese quasi-parallel corpus using analogical associations.

1 Introduction

The recent progress in statistical machine translation and example-based machine translation were driven by the availability of parallel corpora. However, still, acquiring large parallel bilingual corpora is a major bottleneck in developing such machine translation systems in new domains or languages, simply because producing data from scratch is expensive and time-consuming. One of the most important contribution to the development of statistical machine translation was the release of the Europarl Corpus[1], which contained 11 official languages of the European Union in Release V3, now extended to 21 languages in Release v7. In contrast, almost no free or large parallel corpus for Chinese-Japanese is available. In consequence, there is an urgent need for bilingual Chinese-Japanese parallel corpora. Some researchers already addressed the issue and proposed to build such corpora by hand. This is expensive and time-consuming. There are also some automatic approaches to build such corpora. Usually, acquisition of a parallel corpus for the use in language processing tasks typically takes five steps (Koehn [4]): data collection (obtain the raw data from Web by crawlers), document alignment, sentence splitting, segmentation and tokenisation ,finally sentence alignment. Following a recent trend, we propose to construct of a Chinese-Japanese corpus entirely automatically. Attempts at mining parallel text from web using common crawler (Resnik and Smith [8]; Smith et al. [10]) show the feasibility to extract parallel text from Web (dirt and cheap mining). In a variety of recent works, Aker and Gaizauskas [1], Munteanu et al. [7] or Tsunakawa et al. [11] have used comparable corpora to extract parallel sentences or phrases. Other researchers (Lepage and Denoual [6]; Fujita et al. [3]) show how to acquire large collections of paraphrases through generalization and instantiation. Yang et al. [12] investigate the task of acquiring quasi-parallel sentences according to similarities between seed sentences and similarities between analogical clusters. All mentioned works indicate a different solution to build Chinese-Japanese bilingual linguistic resources. By applying analogical techniques, it is more easier and simpler to acquire new quasi-parallel sentences without problems of copyright restriction.

This paper describes the methods and techniques used in our proposed approach to the acquisition of a quasi-parallel Chinese-Japanese corpus. Section 2 describes how to extract analogical clusters from linguistic resources crawled from the Web. Then Section 3 present the core of our our proposed approach. Finally, Section 4 concludes a previous works on our results and draws on future works.

2 Analogical Learning

2.1 Proportional Analogy

Analogical techniques have been applied to several natural language processing tasks. A proportional analogy is a relationship between four objects, noted A: B :: C : D, which reads "A is to B as C is to D". A possible formalization (Lepage [5]) reduces to the counting of number of symbol occurrences and the computation of edit distances. It comes with an efficient algorithm to solve analogical equations between sentences.

$$A : B :: C : D \Longrightarrow \begin{cases} |A|_a - |B|_a = |C|_a - |D|_a, \ \forall a \\ \mathrm{d}(A,B) = \mathrm{d}(C,D) \\ \mathrm{d}(A,C) = \mathrm{d}(B,D) \end{cases}$$

Here, $|A|_a$ stands for the number of occurrences of character a in string A and $d(A,B)$ stands for the edit distance

[1] http://www.statmt.org/europarl/

between strings A and B with only insertion and deletion as edit operations. In this definition, B and C may be exchanged. In this paper, we will use proportional analogy to create rewriting models so as to produce new parallel sentences.

2.2 Creation of Analogical Clusters

We define analogical clusters as sets of pairs of sentences from which any two lines is a proportional analogy. The following cluster of three lines, stand for the following 3 proportional analogies:

A : B :: C : D
A : B :: E : F
C : D :: E : F

The following is an analogical cluster between sentences in Japanese which follows our definition:

紅茶が飲みたい。 : ビールが飲みたい。
紅茶が好きです。 : ビールが好きです。
紅茶は苦手です。 : ビールは苦手です。
紅茶を飲みます。 : ビールを飲みます。

It stands for 6 proportional analogies. In order to obtain analogical clusters, we collect short Japanese and Chinese sentences from the Web using an in-house Web-crawler. We then create all analogical clusters from these sentences in each languages. Table 1 gives the details about the sentences and the number of created analogical clusters. In the experiment, we eliminate sentences containing only numbers and symbols. In the experiment,

# of sentences	collected	filtered	unique
Chinese	810,541	394,035	325,815
Japanese	737,727	446,953	433,292

Table 1: Statistics about crawled short sentences. About half, or more, are kept after filtering

we also eliminate meaningless clusters containing number substitutions or date substitutions. Table 2 shows the details.

	training set	# of obtained clusters	# of sifted clusters
Chinese	17.3M	95,012	78,630
Japanese	20.1M	135,111	31,407

Table 2: Statistics about training set, created analogical clusters and finally sifted clusters

3 Acquisition of New Sentences

3.1 Test Set

We use bilingual texts crawled from a specific Japanese learning website[2] and process data as explained above.

[2] http://jp.hjenglish.com

In this way, we produce 16,246 Chinese-Japanese quasi-parallel sentence pairs as the test set for our experiments. Though there are copyright concerns, we only make use of the sentences we crawled as the seed sentences to feed our system. The new sentences that will be generated can be said to be out of the scope of copyright, and they can be released and made publicly available for this reason.

3.2 Analogical Generation

To generate new sentences based on analogical clusters, we follow Saussure [9] and consider analogical equations as a synchronic operation to produce new forms.

$$wolf : wolves :: leaf : x \Longrightarrow x : leaves$$

Given two forms of a word and only one form of a second word, the fourth missing form is coined by proportional analogy. We apply the same principle to sentences. E.g.,

紅茶が飲みたい。 : ビールが飲みたい。 :: 紅茶が好きです。 : X (1)

The solution of this analogical equation in X is:

X = ビールが好きです。

In this study, we generate new sentences using a set of analogical clusters $\mathscr{C}$. Given an analogical cluster $\mathscr{C}[i] = \{(X_j, Y_j) | j \in [0,1,...,J]\}$, J denotes the number of sentence pairs in cluster $\mathscr{C}[i]$. We make use of analogical clusters as rewriting models to generate new sentences. A line $< X_j : Y_j >$ in $\mathscr{C}[i]$ ($\mathscr{C}[i] \in \mathscr{C}$), assume $[X_j : Y_j :: seed : \mathrm{X}]$ has a solution, we can easily get the new sentence X. In the experiment, we underline generating new sentences using both $A : B :: C : X$ and $B : A :: C : X$.

# of sentences	seed	generated	unique
Chinese	16,246	204,021,052	75,763,878
Japanese	16,246	71,723,057	14,383,579

Table 3: Some statistics about seed sentences, generated sentences and different sentences

New sentences generated from the same seed sentence will be stored in the same file. Since we have a parallel Chinese-Japanese corpus, i.e, a list of (source, target) sentence pairs at our disposal, we apply analogical generation to obtain new sentences using these parallel sentences as seed. By remembering the translation corresponding between seed sentences, we deduce the translation correspondence between generated sentences. Over hundred millions of new sentences are produced, but not all newly generated sentences are valid. We assessed the productivity of our sentences on a sample of 500 seed sentences.

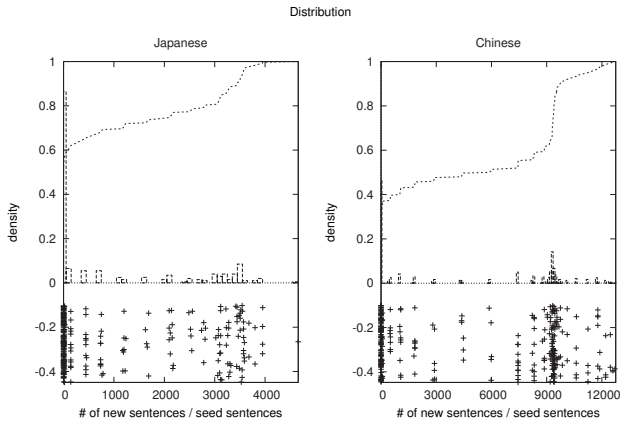

Figure 1: Generation of new sentences in Chinese and Japanese. Each point stands for one seed sentence, the number of new sentences can be generated using all analogical clusters

3.3 Filtering New Sentences

During generation of new sentences, a lot of semantically invalid and grammatically incorrect sentences are produced. The method we use to ensure fluency and adequacy of generated sentences is to eliminate any sentences that contains an N-sequence unseen in the initial corpus. This is conform to the trend of using N-sequences (Doddington [2]) in natural language processing tasks. Table 4 and Table 5 gives statistics on the filtered generated sentences. To assess correctness, we request bilingual speakers to judge the new sentences (see Table 3). The translation relationship correspondence between newly produced sentences depends on the similarity of analogical clusters across the languages. Figure 1 illustrates how different numbers of seed sentences usually generate similar number of new sentences. Each seed sentence can generate thousands of new sentences.

		# of new sentences per seed		
	# of filtered	average	std.deviation	precision
N=4	67,991	4.18	10.95	82%
N=5	17,084	1.05	2.14	95%
N=6	9,756	0.66	0.96	99%
N=7	7,467	0.46	0.66	99%

Table 4: Statistics about filtered generated sentences for Japanese

		# of new sentences per seed		
	# of filtered	average	std.deviation	precision
N=4	59,667	3.67	14.26	90%
N=5	21,011	1.29	2.35	91%
N=6	18,303	1.13	1.08	99%
N=7	17,593	1.08	0.99	99%

Table 5: Statistics about filtered generated sentences for Chinese

# of	Chinese	Japanese
different sentences	5,999	6,620
different clusters	348	343

Table 6: Statistics about sentences and clusters which are found in sentences generation

# of pairs	$sim_{cluster} \geq 0.5$
generated sentence	31,330
cluster	2,643
sifted sentence	11,916

Table 7: Result of quasi-parallel sentences generation

For a total of 16,246 pairs of seed parallel sentences, we obtained 31,330 translation candidate pairs coming from 6,420 pairs of seed sentences, using a filtering threshold of 5.

3.4 Sentence Alignment

Since the new sentences are constructed of seed sentence pairs and analogical clusters, we align the new sentences based on the similarity of seed sentences (sim_s) and the similarity of analogical clusters (sim_c) in both languages. It is possible to compute similarity using lexical weights based on a word-to-word dictionary. Given a sentence pair $<S,T>$, the target sentence T, the source sentence S and a a word alignment between the target word position $i = 0,1,\ldots,I$ in sentence and the source word positions $j = 0,1,\ldots,J$. Similarly to lexicon weights, the similarity between seed sentences, sim_{seed}, can be computed according to following formula:

$$sim_{seed}(T|S) = \prod_{i=1}^{n} \frac{1}{|\{j|(i,j) \in a\}|} \sum_{\forall (i,j) \in a} p(t_i|s_j) \quad (2)$$

Where $p(t_i|s_j)$ is the probability of s_j translates to t_i. We make use of similarity between analogical clusters as one of the weights to deduce the translation relation of newly generated sentence pairs. We also use the Longest Common Sequences (LCS) as Yang et al. [12] proposed to give the automatic scores as the similarity of clusters. In addition to these automatic scores, we asked bilingual speakers to give scores between clusters. Our guidelines are: 1.0 (same), 0.8 (very related), 0.5 (partially related) and 0 (no relation). Based on similarities, we implement the experiments. Table 6 indicates how many different sentences and clusters are found in a single language among these translation candidate pairs. Table 7 shows a more specific experiment result of sifting these newly obtained data. We select all new sentences ($sim_{cluster} \geq 0.5$) as the sentences to construct our quasi-parallel corpus for Chinese-Japanese. As a final result, 11,916 of new quasi-parallel sentences are obtained.

4 Conclusions

This paper introduced a technique for the construction of an open-source Chinese-Japanese quasi-parallel corpus. It uses an expansion filtering technique. Expansion relies on generation by proportional analogy, filtering is done by checking the presence of N-grams in a reference corpus. Future work may focus on finding a way to measure the similarity between analogical clusters and break sentences into phrases to apply the proposed technique to smaller pieces. We have proposed to deduce translation relationship according to similarities between analogical clusters and seed sentences. In order to be able to recognize quasi-parallel sentences, to be able to identify corresponding cluster is a prerequisite. For this, better way of computing similarity between clusters is required.

References

[1] Y. Aker, A. Feng and R. Gaizauskas. Automatic bilingual phrase extraction from comparable corpora. In *COLING 2012, IIT Bombay, Mumbai, India*, 2012.

[2] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research (HLT'02)*, pages 138–145. Morgan Kaufmann Publishers Inc., 2002.

[3] Atsushi Fujita, Pierre Isabelle, and Roland Kuhn. Enlarging paraphrase collections through generalization and instantiation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pages 631–642. Association for Computational Linguistics, 2012.

[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, 2005.

[5] Yves Lepage. Solving analogies on words: an algorithm. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL '98)-Volume 1*, pages 728–734. Association for Computational Linguistics, 1998.

[6] Yves Lepage and Etienne Denoual. Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation. In *Proc. of the 3rd Int. Workshop on Paraphrasing*, pages 57–64, 2005.

[7] Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *HLT-NAACL 2004: Main Proceedings*, pages 265–272, 2004.

[8] Philip Resnik and Noah A. Smith. The Web as a parallel corpus. *Computational Linguistics*, 29(3): 349–380, 2003.

[9] F de Saussure. Cours de linguistique générale. *Paris: Payot.(1st ed. , 1916)*, 1995.

[10] Jason R. Smith, Philipp Koehn, Herve Saint-Amand, Chris Callison-Burch, Magdalena Plamada, and Adam Lopez. Dirt cheap Web-scale parallel text from the common crawl. In *The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria*, 2013.

[11] Takashi Tsunakawa, Naoaki Okazaki, Xiao Liu, and Jun'ichi Tsujii. A Chinese-Japanese lexical machine translation through a pivot language. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(2):9, 2009.

[12] Wei Yang, Hao Wang, and Yves Lepage. Using analogical associations to acquire Chinese-Japanese quasi-parallel sentences. In *The Tenth Symposium on Natural Language Processing (SNLP 2013), Phuket, Thailand*, 2013.