Quasi-Parallel Corpora: Hallucinating Translations for the Chinese–Japanese Language Pair

Yves Lepage
Waseda University
808-0135 Fukuoka-ken, Kitakyûsyû-si, Wakamatu-ku, Hibikino 2-7, Japan
yves.lepage@waseda.jp

Abstract

We show how to address the problem of bilingual data scarcity in machine translation. We propose a method that generates aligned sentences which may be not perfect translations. It consists in 'hallucinating' new sentences which contain small but well-attested variations extracted from unaligned unrelated monolingual data. We conducted various experiments in statistical machine translation between Chinese and Japanese to determine when adding such quasi-parallel data to a basic training corpus leads to increases in translation accuracy as measured by BLEU.

Keywords: Machine translation, Quasi-parallel data, Comparable corpora

1. Introduction

Some languages are well-resourced. This means that tools like segmenters, morphological analysers, syntactic or semantic parsers are available for them. It also means that large amounts of monolingual data are available, usually freely available. Some language pairs are also well-resourced. This means that large amounts of parallel, i.e., well-aligned, data exist for the two languages. Indeed a large number of language pairs are not well-resourced, so that directly building translation systems for these languages is problematic. In this respect, one-shot translation (Johnson et al., 2016) in the framework of neural machine translations raises great expectations. Nevertheless, it is still acknowledged that the lack of aligned or parallel data remains a problem for MT in general.

2. Lack of Parallel Data for Chinese–Japanese

2.1. The Situation

Individually, Chinese and Japanese are relatively well-resourced languages with efficient segmenters, morphological analysers, parsers, etc. However, the language pair itself suffers from a lack of freely available bilingual corpora and this is a problem for machine translation between these two languages.

The BTEC corpus (Takezawa et al., 2002) contains short sentences in the tourism domain, but this corpus is not available for free[1]. The original version contains 160,000 sentences, but it has been extended to more than 1 million. There also exist one large corpus in the scientific and technological domain, used in the MT evaluation campaign WAT, the ASPEC-JC corpus (Nakazawa et al., 2016). Its use requires to sign a license agreement, to participate in the WAT campaign, and to erase data after a one-year term.

2.2. Possible Answers

Different possible solutions to augment the size of parallel corpora have been proposed in the past. They range from the manual creation of data to the automatic extraction of comparable corpora, with attempts at creating bilingual data from monolingual data (Klementiev et al., 2012; Sun et al., 2013; Chu et al., 2013). In statistical machine translation, where the translation table is crucial, directly augmenting the data in the translation table has also been proposed (Luo et al., 2013). All these methods may solve the problem of data scarcity to some extent and lead to increases in BLEU points in different language pairs when used in addition to existing training data.

2.3. The Proposed Answer

The purpose of this paper is to describe a method to create a corpus of aligned sentences, which are translations of one another only up to a certain extent. Because the translation correspondence may not be perfect, we call such a bilingual corpus a quasi-parallel corpus. The similarities and differences between a quasi-parallel corpus and a comparable corpus can be summarised as follows:

Comparable corpus	Quasi-parallel corpus
not exact translations	not exact translations
natural texts	synthetic data
unit: document	unit: sentences
not sentence-aligned	sentence-aligned by design
usually one topic / doc.	any topic

The method consists in 'hallucinating' linguistic data (Irvine and Callison-Burch, 2014), i.e., in creating hopefully parallel, synthetic data from unrelated unaligned monolingual data. However, a certain amount of parallel data as seed data is necessary.

In previous works, we assessed different sets of such 'hallucinated' data by adding them to a training corpus to build an SMT system. This led to variable improvements, as measured by BLEU, ranging from less than half a point on difficult tasks, to several points in other tasks (Wang et al., 2014a), depending on the experimental conditions.

3. Generation of Quasi-Parallel Corpora

3.1. Collecting Variations in Monolingual Data

Figure 1 gives an illustrated overview of the proposed method. The central object in the method is a list of analog-

[1] Approximate cost as of February 2018: 1 yen per sentence.

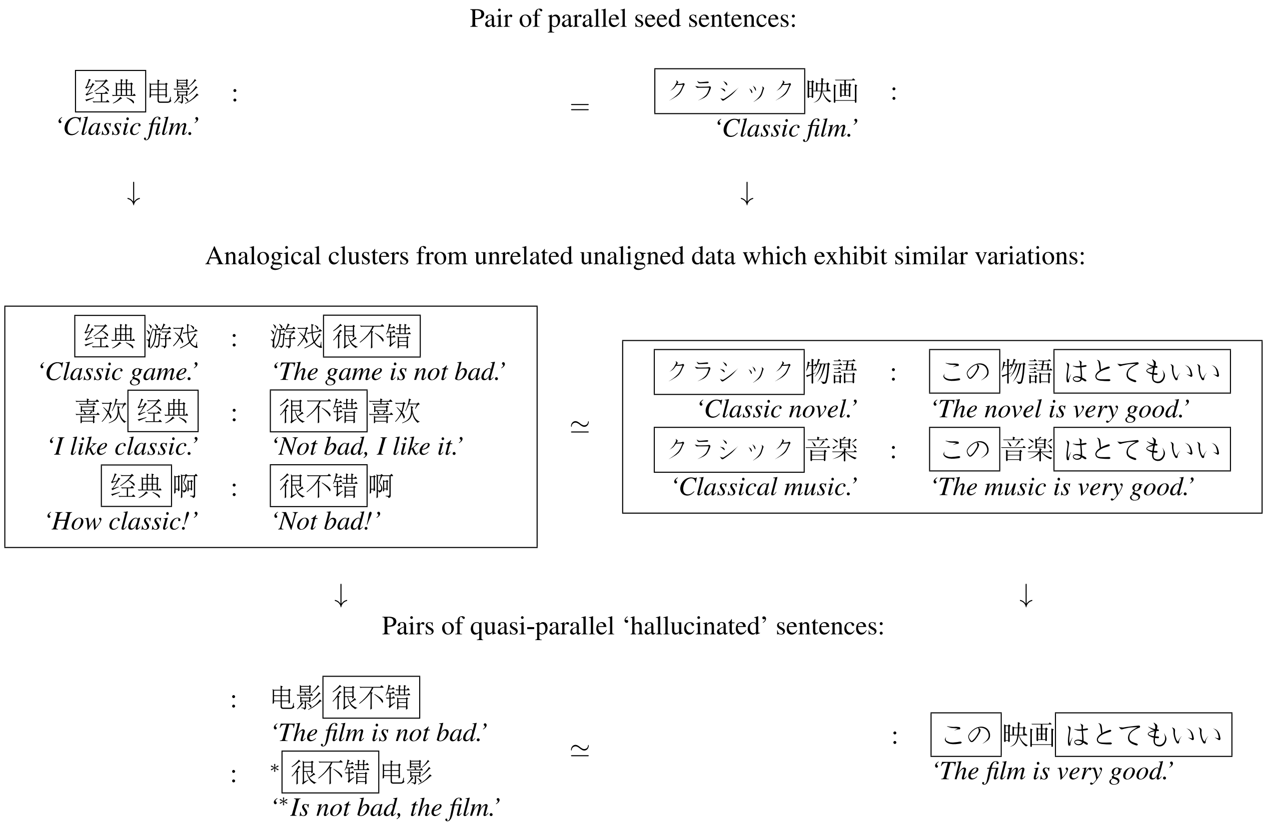

Figure 1: Overview of the generation of hallucinated quasi-parallel translations from parallel seed sentences using analogical clusters produced from unrelated unaligned monolingual data. Chinese on the left, Japanese on the right. The clusters exhibit similar variations so that the sentences obtained from aligned seed sentences can be thought to be almost translations of one another. The variations exhibited by the clusters are framed. The Japanese part shows that the variations may be discontinuous. Notice that the sentences in the analogical clusters are not translations and that the number of sentences in each cluster is different in each language. Notice also that the second hallucinated Chinese sentence is ungrammatical.

ical clusters which exhibit similar variations. In this example, the variation in both languages can be represented as:

$$\boxed{\text{经典}}\,\mathrm{X} : \mathrm{X}\,\boxed{\text{很不错}} \simeq \boxed{\text{クラシック}}\,\mathrm{X} : \boxed{\text{この}}\,\mathrm{X}\,\boxed{\text{はとてもいい}}$$

'Classic X.' *'X is not bad.'* *'Classic X.'* *'This X is very good.'*

Basically, this is an illustration of the principle of translation by analogy introduced in (Nagao, 1984). Finding such configurations requires to perform two tasks: firstly, to collect a relatively large number of small variations in each language, which are well-attested; secondly to be able to show that some of the well-attested monolingual variations in one language correspond to some well-attested variations in the other language.

To complete the first task, we deal with the idea of well-attested series of variations (Yang et al., 2013a; Yang et al., 2013b). Such series of variations, extracted from actual monolingual data, are shown in Figures 2 and 3 for Chinese and Japanese respectively. They are instances of what is called analogical clusters. For details concerning the definition of analogical clusters and the fundamental relation this definition relies on, i.e., proportional analogy, see Appendix 7.

As for implementation, (Fam and Lepage, 2018) describes a set of publicly released tools to automatically output analogical clusters from textual data. The example clusters in Figures 2 and 3 have been obtained using these tools.

3.2. Similarity of Variations Across Languages

In order for the method illustrated in Figure 1 to work, it is necessary to complete the second task mentioned in the previous section, i.e., to be able to show that some of the well-attested monolingual variations in one language correspond to some other well-attested variations in the other language. For that, we use classical ways of comparing bags-of-words across languages.

The computation is performed on the variations exhibited in a cluster. Hence, we compute the differences between the left and the right sides of each cluster in each language and compare these differences by use of Dice coefficients. In order to normalise words across languages, in the case of Chinese and Japanese, we make use of hanzi-kanji conversion tables and dictionaries. The use of translation tables is of course possible. See Appendix 8. for formulae used in estimating the similarity between analogical clusters across two languages.

As shown in the appendix, a reasonably high value of 0.833 is obtained for the two clusters shown in Figure 1.

不值得购买	:	很值得购买
'It's not worth buying.'	:	*'It's worth buying.'*
这个游戏不好玩	:	这个游戏很好玩
'The game is not funny.'	:	*'The game is very funny.'*
画面不好	:	画面很好
'The frame is bad.'	:	*'The frame is very good.'*
小朋友不喜欢	:	小朋友很喜欢
'Children don't like it.'	:	*'Children like it very much.'*
难度不大	:	难度很大
'It's not difficult.'	:	*'It's very difficult.'*
⋮	:	⋮

太好了	:	效果太好了
'It's very good.'	:	*'The effect is very good.'*
非常不错	:	效果非常不错
'It's not bad.'	:	*'The effect is not bad.'*
画面很好	:	画面效果很好
'The frame is very good.'	:	*'Effect of the frame is very good.'*
很炫	:	效果很炫
'It's very cool.'	:	*'The effect is very cool.'*
马马虎虎	:	效果马马虎虎
'It's just so-so.'	:	*'The effect is just so-so.'*
⋮	:	⋮

Figure 2: Two analogical clusters in Chinese. The first one *(top)* illustrates the opposition between negative and affirmative sentences (不 *'not'* replaced by copula 很 *'is'* (etymologically adverb *'very'*)). The second one *(bottom)* illustrates the replacement of unexpressed subjects (expressed in English by the pronoun *'it'*) by the noun 效果 *'effect'*. The framed sentence shows that the same sentence may be found in different analogical clusters.

3.3. Generating Hallucinated Synthetic Data by Application of the Variations

As Figure 1 illustrates, it is possible to apply the variations exhibited in an analogical cluster to any sentence for which it makes sense. The very application of the variations on a sentence is performed by solving equations. E.g., for Figure 1, the equation

$$\boxed{\text{经典}}\text{游戏} : \text{游戏}\boxed{\text{很不错}} :: \boxed{\text{经典}}\text{电影} : x$$

is formed by taking the first line in the Chinese cluster and the Chinese sentence in the pair of aligned sentences at the top of the figure. The solution of this equation is the first Chinese 'hallucinated' sentence: $x = \text{电影}\boxed{\text{很不错}}$.

As all the lines in a cluster are used in turn, it is understandable that the same hallucinated sentence may be generated several times.

3.4. Filtering Out Ill-Formed Sentences

However, as mentioned in the caption of Figure 1 and as is well known with analogy, there is a danger of over-generation, i.e., a risk of creating sentences which are ill-formed, either because they make no sense (ill combinations of characters) or because they are ungrammatical.

早急に対応して下さい。	:	早急に対応して欲しい。
'Please respond as soon as possible.'	:	*'I want you to respond as soon as possible.'*
正式版に戻して下さい。	:	正式版に戻して欲しい。
'Please return to the official version.'	:	*'I want you to return to the official version.'*
元に戻して下さい。	:	元に戻して欲しい。
'Please return to the beginning.'	:	*'I want you to return to the beginning.'*
やめて下さい。	:	やめて欲しい。
'Please stop.'	:	*'I want you to stop.'*
⋮	:	⋮

本当に良かった	:	良かったですね
'It was really good.'	:	*'It was good.'*
本当に酷い	:	酷いですね
'It was so cool.'	:	*'It was cool.'*
本当に嬉しい	:	嬉しいですね
'I am really very happy.'	:	*'I am very happy.'*
⋮	:	⋮

Figure 3: Two analogical clusters in Japanese. The first one *(top)* illustrates the opposition between a request or a wish expressed by 下さい *'Please'* and 欲しい *'I want'* respectively at the end of the sentence. The second one *(bottom)* illustrates the opposition between informal speech on the left and standard speech on the right (suffixation by a copula です and a sentence marker ね). In addition, the sentences on the left include 本当に *'in fact, really, in reality'* at their beginning.

This is the case with the second Chinese hallucinated sentence in Figure 1.

To remedy this problem, based on extensive experiments and comparison of different methods (SVM, language models), we rely on counts of N-sequences to check for naturalness (Doddington, 2002; Lin and Hovy, 2003). The results of our experiments suggest to take a rigid stance and to reject any sentence which contains an N-sequence not attested in a given reference dataset. In other terms, for a sentence to be retained, all of its N-sequences should be attested in the reference dataset (N = 6 for Chinese and 7 for Japanese in our experiments). The method favours precision to the detriment of recall. Indeed manual inspection suggests that a very large amount of valid sentences are actually discarded. However, in experiments where we assessed the quality of the retained sentences, it was judged that 99 % of the sentences are correct in Chinese and Japanese. As for reference dataset, the monolingual data used to collect analogical clusters or the training data to be used in an MT experiment can be used.

4. Assessment with Statistical Machine Translation

In various experiments in SMT conducted over several years in different settings (Wang et al., 2014a; Wang et al., 2014b; Yang et al., 2014; Yang and Lepage, 2014b; Yang and Lepage, 2014a; Yang et al., 2015; Yang et al., 2017), it was shown that the introduction of the small variations

	Training data (# of lines)	Additional data (# of lines)	(percent)	Baseline	Increase (rounded BLEU points)
PolTAL 2014 / IPSJ 2017	subtitles_1 110,000	seeds = subtitles_1 clusters = Web news_1 75,000	$\boxed{(+68\%)}$	$11 \sim 13$	$\boxed{+6}$
PIC 2014	subtitles_2 120,000	seeds = 10% of subtitles_2 clusters = Web news_2 10,000	(+8.3%)	$17 \sim 20$	$+2 \sim 4$
WAT 2014	ASPEC corpus 670,000	seeds = 1/6 of ASPEC (length $\leq$ 30 chars) clusters = Web news_1 35,000	(+5%)	(Moses 1.0) $23 \sim 29$	$+2 \sim 3$
Additional exp. 2017	ASPEC corpus 670,000	seeds = 1/6 of ASPEC (length $\leq$ 30 chars) clusters = ASPEC 2,800	$\boxed{(+0.3\%)}$	(Moses 2.1) $30 \sim 37$	$\boxed{+0}$

Table 1: Synthesis in numbers of several experiments in using quasi-parallel corpora for SMT. Subtitles_1 and Subtitles_2 are different excerpts from the OpenSubtitles corpus (Tiedemann, 2009). Web news_1 and Web news_2 are two in-house datasets browsed from various news sites in Chinese and Japanese. Larger improvements are obtained when the training corpus and the quasi-parallel corpus are from different domains and when the quasi-parallel corpus is large relatively to the training corpus (compare framed values on first and last lines).

created by the proposed method of adding a quasi-parallel corpus to the training data explained above, increases the size of the translation tables and that the new phrases are actually used and may contribute to translation accuracy. A synthesis of the results obtained over the years is given in Table 1.

The overall results are mitigated. The improvements in translation accuracy as measured by BLEU vary from large positive values to smaller and less encouraging values. Also, in experiments reduplicated with different versions of the Moses engine, versions 1.0 and 2.1, it was observed that the upgrade of the Moses engine made up for the increases brought by the method on the older version.

Notwithstanding the various improvements in BLEU scores, two main lessons can be drawn from the SMT experiments.

Firstly, several experiments tend to show that the quality of the alignment of the produced sentences is not so crucial. What seems to be crucial is the grammaticality of the sentences produced. For that, different configurations and various methods have been tested so as to automatically ensure a very high level of grammaticality or semantic correctness. The N-sequence filtering method was found to be the most effective technique to filter out ill-formed sentences, despite a very low recall.

Secondly, the relationship or rather the absence of relation between the basic training data and the monolingual data seems to be important. Monolingual data from the same domain or the same collection of texts do not seem to conduct to significant improvements. Thorough experiments still need to be conducted to confirm this impression, but it seems that variations from the general language, are necessary to bring improvements in translation accuracy. Relatively to this, the larger the quasi-parallel corpus added to the training corpus, the better.

5. Conclusions

The method described in this paper to produce a quasi-parallel corpus relies on the application of a large number of small well-attested variations on a relatively small number of parallel seed sentences. As SMT is concerned, these small variations are captured in the translation table and, if such small variations appear in the test set, the test set may be better translated. This is shown by the fact that a larger number of the phrases generated from the quasi-parallel corpus are indeed used to translate the test set, in comparison to a baseline system trained without the quasi-parallel corpus.

What seems important for the method to work is the grammatical quality of the generated sentences, while, relatively, the quality of the correspondence between the clusters may not be so important. It seems that the best configuration is a configuration where the monolingual data for the extraction of analogical clusters is varied enough so as to offer useful variations and where these monolingual data are different from those found in the training data, i.e., new variations can be found. Consequently, the positive effect of the quasi-parallel corpus may be thought as the effect of providing variations found in the general usage of the languages to be translated.

6. Acknowledgements and Thanks

A large part of the work reported here was supported by a JSPS Grant, Number 15K00317 (Kakenhi C), entitled "Language productivity: efficient extraction of productive analogical clusters and their evaluation using statistical machine translation."

Special thanks to YANG Wei (former PhD student) who, as a research assistant, was instrumental and contributed largely to the implementation of the method and to the obtention of the results described and reported here.

Thanks to WANG Hao (PhD student), SHEN Hanfei and GAO Mengru (former master students) for collecting data, writing programs and running experiments.

7. Appendix: Definition of Analogical Clusters

An analogical cluster is defined in the following way, where the s's stand for sentences, i.e., strings of characters (computation in strings of words is also possible):

$$\begin{array}{c} s_1^1 : s_1^2 \\ s_2^1 : s_2^2 \\ \vdots \ \ \vdots \\ s_n^1 : s_n^2 \end{array} \quad \stackrel{\Delta}{\Longleftrightarrow} \quad \begin{array}{r} \forall (i,j) \in \{1,\ldots,n\}^2, \\ s_i^1 : s_i^2 :: s_j^1 : s_j^2 \end{array} \tag{1}$$

In this definition, it is understandable that the underlying relation between four sentences, noted by semi-colons and double semi-colons as $s_i^1 : s_i^2 :: s_j^1 : s_j^2$, is the most important notion. This notion is that of proportional analogy, for which we adopt the characterisation introduced in (Lepage, 1998; Lepage, 2003):

$$A : B :: C : D \ \Rightarrow \begin{cases} |A|_a - |B|_a = |C|_a - |D|_a, \forall a \\ d(A,B) = d(C,D) \\ d(A,C) = d(B,D) \end{cases} \tag{2}$$

where $d(A,B)$ is the distance between two strings A and B and $|A|_a$ stands for the number of occurrences of character a in string A.

In order to make Characterisation (2) operational, we read it in the other direction, i.e., we assume that an analogy holds when the constraints on distance and character counts are met.

8. Appendix: Computation of Analogical Cluster Similarity Across Two Languages

For simplicity, we compare analogical clusters across languages by first extracting the differences in words on their left and right sides and then compare two analogical clusters in two different languages by taking the mean of the Dice coefficients for the differences on each of their sides. This is expressed by Formula (3).

$$\begin{aligned} &\text{Sim}((L_{zh} : R_{zh}), (L_{ja} : R_{ja})) = \\ &\qquad \frac{1}{2}(\text{Dice}(L_{zh}, L_{ja}) + \text{Dice}(R_{zh}, R_{ja})) \end{aligned} \tag{3}$$

We repeat the formula for the Dice coefficient ($|S|$ stands for the cardinality of a set S):

$$\text{Dice}(S_{zh}, S_{ja}) = \frac{2 \times |S_{zh} \cap S_{ja}|}{|S_{zh}| + |S_{ja}|} \tag{4}$$

To be able to compute the intersection between two sets of words in two different languages, Chinese and Japanese, we normalise the words in one language in the other language by making use of kanji-hanzi conversion, dictionaries, translation tables, etc. Nowadays we should consider bilingual word vector representations.

As an illustration, for the clusters in Figure 1, knowing from some dictionary or translation table that 经典 = クラシック, 很 = とても and 不错 = いい, we perform the following computation:

$$\begin{aligned} &\text{Sim}((L_{zh} : R_{zh}), (L_{ja} : R_{ja})) \\ &\quad = \frac{1}{2}\left(\frac{2 \times |\{\text{经典} = \text{クラシック}\}|}{|\{\text{经典}\}| + |\{\text{クラシック}\}|}\right. \\ &\qquad \left. + \frac{2 \times |\{\text{很} = \text{とても}, \text{不错} = \text{いい}\}|}{|\{\text{很}, \text{不错}\}| + |\{\text{この}, \text{は}, \text{とても}, \text{いい}\}|}\right) \\ &\quad = \frac{1}{2}\left(\frac{2 \times 1}{1+1} + \frac{2 \times 2}{2+4}\right) \\ &\quad = \frac{1}{2}\left(1 + \frac{2}{3}\right) \\ &\quad = 0.833 \end{aligned}$$

because the left and right parts of the variations in each of the Chinese and Japanese clusters are

$$(L_{zh} : R_{zh}) = (\{\text{经典}\} : \{\text{很}, \text{不错}\})$$

and

$$\begin{aligned} &(L_{ja} : R_{ja})) = \\ &\qquad (\{\text{クラシック}\} : \{\text{この}, \text{は}, \text{とても}, \text{いい}\}) \end{aligned}$$

respectively.

As the values range from 0 to 1, with higher values showing greater similarity, a value of 0.833 can be interpreted as a high similarity for the variations exhibited by the two clusters.

9. Bibliographical References

Chu, C., Nakazawa, T., and Kurohashi, S. (2013). Chinese–Japanese parallel sentence extraction from quasi–comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 34–42, Aug.

Doddington, G. (2002). Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, San Diego, March. Morgan Kaufmann Publishers Inc.

Fam, R. and Lepage, Y. (2018). Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan, May.

Irvine, A. and Callison-Burch, C. (2014). Hallucinating phrase translation for low resource MT. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 160–170, Ann Arbor, Michigan, June. Association for computational linguistics.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Klementiev, A., Irvine, A., Callison-Burch, C., and Yarowsky, D. (2012). Toward statistical machine translation without parallel corpora. In *Proceedings of the*

13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12, pages 130–140, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lepage, Y. (1998). Solving analogies on words: an algorithm. In *Proceedings of COLING-ACL'98*, volume I, pages 728–735, Montréal, August.

Lepage, Y. (2003). *De l'analogie rendant compte de la commutation en linguistique (Of that kind of analogies capturing linguistic commutation)*. Habilitation thesis, Joseph Fourier Grenoble University, May.

Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL 2003*, pages 150–157, Edmonton, May.

Luo, J., Max, A., and Lepage, Y. (2013). Using the productivity of language is rewarding for small data: Populating SMT phrase table by analogy. In Zygmunt Vetulani, editor, *Proceedings of the 6th Language & Technology Conference (LTC'13)*, pages 147–151, Poznań, December. Fundacja uniwersytetu im. Adama Mickiewicza.

Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. *Artificial & Human Intelligence*, pages 173–180.

Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). Aspec: Asian scientific paper excerpt corpus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2204–2208, Portorož, Slovenia, may. European Language Resources Association (ELRA).

Sun, J., Qing, Y., and Lepage, Y. (2013). An iterative method to identify parallel sentences from non-parallel corpora. In Zygmunt Vetulani, editor, *Proceedings of the 6th Language & Technology Conference (LTC'13)*, pages 238–242, Poznań, December. Fundacja uniwersytetu im. Adama Mickiewicza.

Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S. (2002). Toward a broad coverage bilingual corpus for speech translation of travel conversation in the real world. In *Proceedings of LREC 2002*, pages 147–152, Las Palmas, May.

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, et al., editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Wang, H., Yang, W., and Lepage, Y. (2014a). Improved Chinese-Japanese phrase-based MT quality using extended quasi-parallel corpus. In Yinglin Wang, et al., editors, *Proceedings of the 2014 IEEE International Conference on Progress in Informatics and Computing (PIC 2014)*, pages 6–10. IEEE Computer Society Press, May.

Wang, H., Yang, W., and Lepage, Y. (2014b). Sentence generation by analogy: towards the construction of a quasi-parallel corpus for Chinese-Japanese. In *Proceedings of the 20th Annual Meeting of the Japanese Association for Natural Language Processing*, pages 900–903, Sapporo, March.

Yang, W. and Lepage, Y. (2014a). Consistent improvement in translation quality of Chinese–Japanese technical texts by adding additional quasi-parallel training data. In *Proceedings of the First Workshop on Asian Translation (WAT)*, pages 69–76, October.

Yang, W. and Lepage, Y. (2014b). Inflating a training corpus for SMT by using unrelated unaligned monolingual data. In Adam Przepiórkowski et al., editors, *Advances in Natural Language Processing: Proceedings of the 9th conference on language processing (PolTAL 2014)*, volume LNAI 8686, pages 236–248, Warsaw, Poland, September. Springer.

Yang, W., Wang, H., and Lepage, Y. (2013a). Automatic acquisition of rewriting models for the generation of quasi-parallel corpus. In Zygmunt Vetulani, editor, *Proceedings of the 6th Language & Technology Conference (LTC'13)*, pages 409–413, Poznań, December. Fundacja uniwersytetu im. Adama Mickiewicza.

Yang, W., Wang, H., and Lepage, Y. (2013b). Using analogical associations to acquire Chinese-Japanese quasi-parallel sentences. In *Proceedings of the tenth symposium on natural language processing (SNLP2013)*, pages 86–93, Phuket, Thailand, October.

Yang, W., Wang, H., and Lepage, Y. (2014). Deduction of translation relations between new short sentences in Chinese and Japanese using analogical associations. *International Journal of Advanced Intelligence*, 6(1):13–34.

Yang, W., Zhao, Z., and Lepage, Y. (2015). Inflating training data for statistical machine translation using unaligned monolingual data. In *Proceedings of the 21th Annual Meeting of the Japanese Association for Natural Language Processing*, pages 1016–1019, Kyoto, March.

Yang, W., Shen, H., and Lepage, Y. (2017). Inflating a small parallel corpus into a large quasi-parallel corpus using monolingual data for Chinese–Japanese machine translation. *Journal of Information Processing*, 25:88–99.