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

Yves Lepage
Waseda University
808-0135 Fukuoka-ken, Kitakyūshū-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 translation 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. 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:

<table>
<thead>
<tr>
<th>Comparable corpus</th>
<th>Quasi-parallel corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>not exact translations</td>
<td>not exact translations</td>
</tr>
<tr>
<td>natural texts</td>
<td>synthetic data</td>
</tr>
<tr>
<td>unit: document</td>
<td>unit: sentences</td>
</tr>
<tr>
<td>not sentence-aligned</td>
<td>sentence-aligned by design</td>
</tr>
<tr>
<td>usually one topic / doc.</td>
<td>any topic</td>
</tr>
</tbody>
</table>

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-
Pair of parallel seed sentences:

经典电影 : 'Classic film.'
经典电影 : 'Classic film.'

↓

Analogical clusters from unrelated unaligned data which exhibit similar variations:

经典游戏 : 游戏很不错
‘Classic game.’ 'The game is not bad.'
喜欢经典 : 'I like classic.'
‘Not bad, I like it.’
经典啊 : ‘How classic!’
‘Not bad!’

↓

Pairs of quasi-parallel ‘hallucinated’ sentences:

电影很不错
电影很不错
‘The film is not bad.’
“Is not bad, the film.”

≈

この映画はとてもいい
‘The film is very good.’

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.

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.
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

\[
\text{经典游戏：游戏 (很不错) } \quad \text{经典电影：} x
\]

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: \( x = \text{电影 (很不错) } \).

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.

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

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 (をです and a sentence marker な)). 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
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):

\[
\begin{align*}
\forall (i,j) \in \{1, \ldots, n\}^2, \quad s_i^1 : s_i^2 :: s_j^1 : s_j^2
\end{align*}
\]

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).

\[
\text{Sim}((L_{zh} : R_{zh}), (L_{ja} : R_{ja})) = \frac{1}{2} (\text{Dice}(L_{zh}, L_{ja}) + \text{Dice}(R_{zh}, R_{ja}))
\]

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

\[
\text{Dice}(S_{zh}, S_{ja}) = \frac{2 \times |S_{zh} \cap S_{ja}|}{|S_{zh}| + |S_{ja}|}
\]

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, それでも and 不错 = いい, we perform the following computation:

\[
\text{Sim}((L_{zh} : R_{zh}), (L_{ja} : R_{ja})) = \frac{1}{2} \left( \frac{2 \times |\text{经典 = クラシック}|}{|\text{经典}| + |\text{クラスック}|} + \frac{2 \times |\text{それでも,不错 = いい}|}{|\text{それとも,不错}| + |\text{これ, とても,いい}|} \right)
\]

\[
= \frac{1}{2} \left( \frac{2 \times 1 \times 2}{1 + 1 + 2 + 4} \right) = \frac{1}{2} \left( \frac{1 + 2}{3} \right) = 0.833
\]

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,不错}\}

and

(L_{ja} : R_{ja}) = \{\text{クラスック,これ,とても,いい}\}

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


