Improved Chinese-Japanese Phrase-based MT Quality Using an Extended Quasi-parallel Corpus

Hao Wang
School of Computer Engineering and Technology
Shanghai University
Shanghai, China
Email: oko_ips@ruri.waseda.jp

Wei Yang and Yves Lepage
Graduate School of Information, Production and Systems
Waseda University
Kitakyushu, Japan
Email: {kevinyoogi@akane.; yves.lepage@}waseda.jp

***Abstract*—State-of-the-art phrase-based machine translation (MT) systems usually demand large parallel corpora in the step of training. The quality and the quantity of the training data exert a direct influence on the performance of such translation systems. The lack of open-source bilingual corpora for a particular language pair results in lower translation scores reported for such a language pair. This is the case of Chinese-Japanese. In this paper, we propose to build an extension of an initial parallel corpus in the form of quasi-parallel sentences, instead of adding new parallel sentences. The extension of the initial corpus is obtained by using monolingual analogical associations. Our experiments show that the use of such quasi-parallel corpora improves the performance of Chinese-Japanese translation systems.**

***Keywords*—machine translation; analogy; paraphrasing; quasi-parallel data**

I. Introduction

Recently, phrase-based statistical machine translation (MT) systems [1] have achieved great success in translation quality, but there still exist two major problems for their applications to any language pair: firstly, unknown words are a major problem in such a data-oriented approach. A phrase-based statistical MT system learns its vocabulary from a training corpus and cannot anticipate nor create new words. Secondly, state-of-the-art MT systems demand large parallel corpora to learn their translation knowledge from. The quality and the quantity of the training data exert a direct influence on the performance of translation systems. The increase in available parallel corpora still does not meet the demand for MT, especially in language pairs like Chinese-Japanese, which is still under-resourced. Collecting more and more larger parallel corpora could boost coverage of vocabulary and solve the corpus shortage, but the process of acquiring large parallel bilingual corpora is still expensive and time-consuming. For some languages, it is not so difficult, but for the particular language-pair Chinese-Japanese, the lack of open-source bilingual parallel corpora is still crying.

Since [2] firstly attempted to exploit comparable corpora to improve statistical machine translation (SMT), even non-parallel corpora are applied into helping training MT system [3]. Improving the quality of MT system with optimization of training corpus then came to the attention of researchers to solve these problems. [4] proved that instead of collecting more parallel corpora, it is possible to improve the SMT performance by exploiting full the potential of existing parallel corpora. [5] improved SMT by using monolingually-derived paraphrases. [6] showed that when higher quality of example-based paraphrases are used, the performance improves.

Corpus-based analogical techniques have been widely applied to several fields of Natural Language Processing. [7] pioneered applying analogy to machine translation. [8] investigated transliteration of English proper names into Chinese using analogical methods. Following a recent trend, we propose to build a system to construct Chinese-Japanese quasi-parallel corpus entirely automatically using analogical techniques on the basis of works like [9]. The difference between a parallel corpus and a quasi-parallel corpus is that a quasi-parallel corpus contains sentences that are translations to each other to a certain extent estimated by some similarity scores. In fact, after rewriting using analogy associations, a sentence pair in a quasi-parallel corpus usually shares the same meaning corresponding to the sentence pairs in the parallel corpus, but slight changes in meaning may also happen. The idea is similar to paraphrasing [10], but is less constrained. Analogy can create near meaning, paraphrases cannot. New quasi-parallel sentence pairs are often composed of different expressions, similar paraphrases or even exhibit changed syntactic structures. In this paper, we make use of an extended quasi-parallel corpus built from initial parallel data. Figure 1 depicts the detailed working process of the system to construct a quasi-parallel corpus. A small initial parallel corpus is required to drive the system. Both bilingual parallel resources and monolingual short sentential resources are collected from the Web using several in-house crawlers designed by ourselves. Our system learns analogical knowledge from monolingual corpora and makes use of this knowledge as rewriting models to produce new sentences using analogy. We could generate millions of candidate sentences to build new quasi-parallel sentence pairs. New generated sentences in Chinese and Japanese are aligned to produce new quasi-parallel sentence pairs on the basis of criteria of similarity described below.

The remainder of this paper is organized as followed. In Section 2, related work in data collection is presented. Section 3 describes the notions and applications of analogy, e.g., analogical clustering and analogical rewriting. In Section 4, we report on experiments on the performance of our improved MT system. We describe the evaluation carried out to test and show better evaluation scores. In Section 5, finally we comment the results obtained, draw some conclusions and point out possible future lines of work.

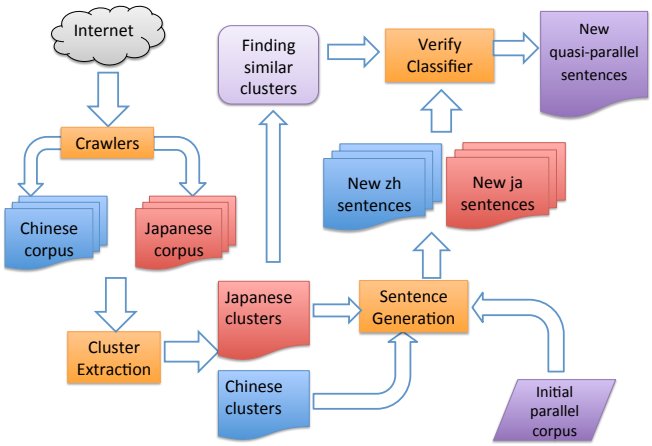

Fig. 1. Flowchart of the construction system of quasi-parallel bilingual corpus for Chinese and Japanese

II. Collection of Linguistic Resources

A. Collection of Monolingual Resources

The basic reason for collecting data from the Web is that these resources are rich in linguistic features. Since our research is aimed at surveying the practicability of constructing a quasi-parallel corpus using analogical associations based on short sentences, we mainly investigate short sentences with less than 30 characters in size. In our experiments, a mass of Chinese and Japanese monolingual sentences have been collected from the Web by our in-house crawlers over one year. We collected data mainly from the following websites: Yahoo China[1], Yahoo China News[2], douban[3] for Chinese and Yahoo! JAPAN[4], Mainichi Japan[5], Rakuten Japan[6] for Japanese. The

TABLE I. Sizes of crawled monolingual raw data, cleaned data and filtered data in Chinese and Japanese.

Size of	Chinese	Japanese
raw data	469.1M	594.2M
cleaned data	109.6M	89.2M
filtered data	17.3M	21.4M

raw data contain undesirable characters, strange symbols and traditional Chinese characters. Cleaning keeps sentences that only contain characters that belong to the Simplified Chinese or Japanese character sets. It should be said that this preprocessing with manual check is complex and time-consuming. Since we are only interested in short sentences, we filter out long sentences. Table I gives the details about the features of the data.

B. Collection of Parallel Resources

Earlier, to collect parallel sentences, printed resources were used to collect parallel sentences as they were readily available. With the increasing chances of accessing to digitally stored texts via the Internet, it is much easier to build a corpus, e.g., the Google N-gram corpus [11]. [12] were the first to present a methodology for building aligned multilingual corpora from movie subtitles. Now the popular subtitle formats are based on time. SubStation Alpha (SSA) is a subtitle file format that allows for more advanced subtitles than the traditional SubRip format (SRT file extension) and other formats. Table II shows

TABLE II. Subtitle file format of SSA, in the table, the bold sentences are parallel.

[Events] Format: Layer, Start, End, Style, Name, MarginL, MarginR, MarginV, Effect, Text
.....
Dialogue: **0,0:19:07.40,0:19:11.12,zh,,0000,0000,0000,, 把纪念碑拆掉然后用来修净水设施**
Dialogue: 0,0:19:11.28,0:19:13.05,zh,,0000,0000,0000,, 什么？拆掉?
.....
Dialogue: **0,0:19:07.40,0:19:11.12,ja,,0000,0000,0000,,モニュメントを解体して，それで直してください**
Dialogue: 0,0:19:11.28,0:19:11.91,ja,,0000,0000,0000,,え
Dialogue: 0,0:19:12.03,0:19:13.05,ja,,0000,0000,0000,,解体

TABLE III. Statistics for the initial parallel corpus built.

	Chinese	Japanese
# of files (.ass)	352	352
size of parallel sentences	26.1M	34.4M
# of parallel sentences (unique)	129,787	129,787
avg. len(w)	7.86	7.03
std.dev. (w)	3.05	2.71

a short example of Japanese subtitles and their Chinese correspondences under this format. We collected these data from the following websites: *Subscene.com* and *Opensubtitles.org*. Each text piece consists of one or two short sentences shown on the screen nearly every second both for Chinese and Japanese. The readers have only a limited time to perceive and understand a given subtitle, so the sentences are usually short and simple. Table III gives the statistics on the Chinese-Japanese initial parallel corpus we built.

III. Analogical Reconstructing

A. Proportional Analogy

A proportional analogy is a structural relationship between four objects, noted $A : B :: C : D$, it reads 'A is to B as C is to D'. An example to is *wolf* : *wolves* :: *leaf* : *leaves*. [13] proposes an effective formalization to solve analogical equations between strings of symbols. It is given below.

$$A : B :: C : D \Leftrightarrow \begin{cases} |A|_a + |D|_a = |C|_a + |B|_a, \forall a \\ d(A,B) = d(C,D) \\ d(A,C) = d(B,D) \end{cases} \quad (1)$$

In this formalization, A, B, C and D are strings of symbols and can be words, chunks, N-grams or sentences written in any writing system. Only insertion and deletion are considered as edit operations here. $|A|_a$ stands for the number of occurrences of character a in string A. We can also exchange B and C. By reducing the formalization to the counting of number of symbol occurrences and the computation of edit distances, it becomes possible more easily to find analogical association between sentences.

[1]http://www.yahoo.cn/, 2013-08-01
[2]http://news.yahoo.cn/, 2013-08-01
[3]http:/www.douban.com/, 2013-08-01
[4]http://yahoo.co.jp/, 2013-08-01
[5]http://mainichi.jp/, 2013-08-01
[6]http://www.rakuten.co.jp/, 2013-08-01

B. Analogical Clusters

We define analogical clusters as sets of sentence pairs like the following three lines:

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

These three lines represent the set of three analogies shown below:

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

This set is obtained by forming an analogy by taking any two lines from the analogical cluster. The following is an actual 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。
紅茶は苦手です。 : ビールは苦手です。

Some examples in English for analogical clusters are also given in Figure 2.

In order to obtain analogical clusters, we collect short Japanese and Chinese sentences from the Web using an in-house Web-crawler. In our experiments, we eliminate sentences containing only numbers and symbols, and remove meaningless clusters containing number substitutions or date substitutions. Table IV shows the details.

TABLE IV. STATISTICS ABOUT TRAINING SET, CREATED ANALOGICAL CLUSTERS AND FINALLY SIFTED CLUSTERS

	Size of training set	# of obtained clusters	# of sifted clusters
Chinese	17.3M	92.8M	76.7M
Japanese	21.4M	132.0M	30.6M

C. Analogical Generation

We follow [14] and consider analogical equations as a synchronic operation to produce new forms, e.g.:

$$wolf : wolves :: leaf : x \Rightarrow x = leaves$$

By using this operation, we implement sentence generation based on analogical clusters. We make use of analogical clusters as rewriting models to generate new sentences. Given an analogical cluster $\mathscr{C}[i] = \{(X_j, Y_j) | j \in [0,1,...,J]\}$, where J denotes the number of sentence pairs in the cluster. Line $< X_j : Y_j >$ in $\mathscr{C}[i]$ ($\mathscr{C}[i] \in \mathscr{C}$) in conjunction with a seed sentence gives two analogical equations: $[X_j : Y_j :: seed : X]$ and $[X_j : Y_j :: seed : Y]$. We collect all solutions X and Y when they exist. In our experiments, each seed sentence can generate hundreds of thousands of new sentences. Figure 2 illustrates the procedure of using created rewriting models to produce new sentences.

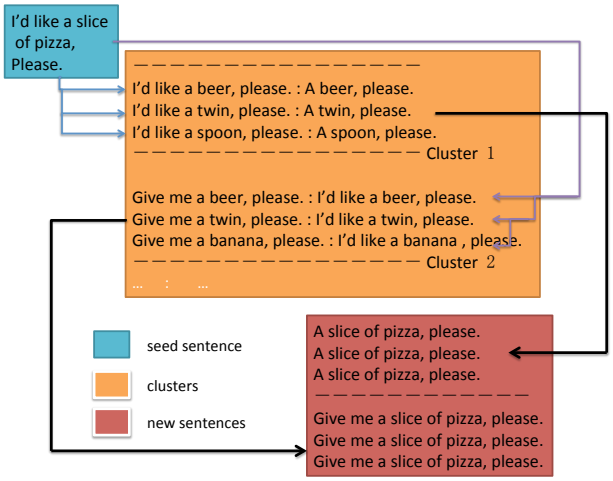

Fig. 2. An example of sentence generation using analogical cluster in English

D. Filtering New Sentences

Given the probability of inconsistency between sentences in Chinese and Japanese, it is essential to restrict new generated sentences by filtering them based on some criteria. 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 sentence that contains an N-sequence unseen in the initial corpus. This is conform to the trend of using N-sequences [15] in natural language processing tasks.

E. Finding Similar Clusters

In all generality, the similarity between seed sentences and the similarity between used clusters should be used to judge the similarity between new sentences. Here, we only use the similarity between analogical clusters because we already know that the seed sentences are parallel. Finding the similar clusters across languages reduces to compute the similarity of the differences between the left parts and the right parts of the Chinese and Japanese clusters. We make use of

- the EDR dictionary[7];
- a traditional-simplified Chinese conversion table (Unicode Data-Traditional-Simplified-Variant)[8]);
- a Kanji-Hanzi Conversion Table[9].

to obtain word translations. Assume that we have two word bags A and B, we define the similarity between clusters using the Dice formula:

$$Sim(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|} \quad (2)$$

We use the Longest Common Sequences (LCS) as proposed in [9] to give the automatic scores as the similarity of clusters. For example, the similarity between two clusters A (means

[7]The EDR Electronic Dictionary: National Institute of Information and Communication Technology (NICT) URL: http://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html,2013-02-01

[8]http://www.unicode.org/Public/, 2013-02-01

[9]http://www.kishugiken.co.jp/cn/, 2013-02-01

TABLE V. EXAMPLES OF CORRESPONDING CLUSTERS WITH HIGH SIMILARITY SCORES IN OUR RESULT.

ja⇔zh	Changing patterns (ja)	Changing patterns (zh)	Similarity
EDR dictionary	事業：友達	事业: 朋友	1.0
+	A: **映画：初恋**	B:**电影: 的 初恋**	0.833
kanji-hanzi conv.	君：あなた	你: 他	1.0
	ε：者	的: 人	0.5
	ε： 彼女	ε：她	1.0

"movie"→"first love") and B ("movie"→"someone's first love") in the TableV is calculated as follows:

$$Sim(A,B) = \frac{1}{2}\left(\frac{2\times|映画|}{|电影|+|映画|} + \frac{2\times|初恋|}{|初恋,的|+|初恋|}\right) \quad (3)$$

$$= \frac{1}{2}\left(\frac{2\times|1|}{|2|} + \frac{2\times|1|}{|3|}\right) = 0.833 \quad (4)$$

In our experiment, we select all new sentence pairs with sim≦0.5 to construct our quasi-parallel corpus for Chinese-Japanese.

IV. EXPERIMENTAL CONTEXT

We obtained a quasi-parallel corpus from an initial parallel corpus, and evaluated Chinese-Japanese MT performance by appending the new quasi-parallel sentences to the training data. We make use of built subtitle corpus for training and a sample of the Japanese-English-Chinese Basic Sentence Data (JEC corpus) [10] for tuning and testing. In our experiment, segmentation of sentences in words is not required in the steps of data preprocessing and finding similar analogical clusters. For statistical machine translation, word segmentation is required, we make use of two toolkits, one for each language:

- **Mecab**: part-of-speech tagger and morphological analyzer for Japanese[11] (Kyoto University & NTT Communication Science Laboratories);
- **Urheen**: Chinese lexical analysis toolkit (National Laboratory of Pattern Recognition, Institute of Automation of the Chinese Academy of Sciences).

A. Experimental Setup

The most widely used state-of-the-art tools to obtain phrase-based machine translation tables now is GIZA++ [16], which trains the IBM models [17] and the HMM introduced by [18], in combination with the Moses toolkit [1]. We use the Moses decoder and MERT (Minimum Error Rate Training) to tune the parameters of the translation tables, and the SRI Language Modeling toolkit [19] for the target language model.

- Baseline configuration: Using all the 120K parallel sentences in subtitles corpus as training data (containing 5K sentences in subset in JEC corpus).
- Tuning: Using another 500 sentences in JEC corpus that are not included in the training set.
- Test: Using another 100 sentences in JEC corpus.
- Extended configuration: same as baseline, but we add 10K of quasi-parallel sentences to the training data. The total number of the sentences for training is thus 130K.

The reason why we evaluate on JEC corpus is that most of the sentences in these data are very simple and colloquial, which is similar to the sentences in the subtitle corpus.

B. Experiment Results

As for the evaluation of translations, we follow standard machine translation evaluation procedure using standard automatic evaluation metrics. We test the statistical significance of the difference between two translation systems trained using different training sets on the same test set. See Table VI for the result.

TABLE VI. EVALUTION RESULTS FOR CHINESE-JAPANESE TRANSLATION ACROSS TWO SYSTEMS, BOLD SCORES ARE SIGNIFICANTLY BETTER.

	Exper.	WER	BLEU	NIST
ja-zh	baseline	0.7471	17.19	3.6621
	+extended	**0.7280**	**20.83**	**3.8716**
zh-ja	baseline	0.7348	19.51	3.5126
	+extended	**0.6643**	**22.11**	**3.6901**

[20] introduced a bootstrap resampling method to compute the statistical significance of test results. We test the hypothesis that the system trained with our extended training set is better than the baseline. The results show a 95% confidence that the extended system is better than the baseline (see Table VII). Except for the BLEU scores on the first line (source-target: from Chinese to Japanese), the p-value shows an acceptable significance level (p-value=0.05).

TABLE VII. EVALUATION OF THE STATISTICAL SIGNIFICANCE OF TEST RESULT IN BOTH DIRECTION, THE NUMBER SHOWN IN TABLE IS THE P-VALUE SUPPORT THOSE HYPOTHESES

source⇒ target	p-value (BLEU)	p-value (NIST)
zh-ja	0.065	0.049
ja-zh	0.005	0.023

V. CONCLUSION

Lack of sufficient linguistic resources for Chinese and Japanese is currently one of the major bottleneck in further advancement of automated translation for this language pair. This paper introduced the efforts and experiments we made to improve the quality of statistical machine translation system using an open-source Chinese-Japanese quasi-parallel corpus instead of parallel corpus. Its main purpose was to propose, analyse and evaluate a novel method by which an extended quasi-parallel corpus was shown to compensate for this shortage of linguistic resources. The method leads to significant improvement for an under-resourced language-pair.

The method consists in an expansion-filtering technique. Expansion relies on generation by proportional analogy; filtering is done by checking the presence of N-grams in a reference

[10] (2011/7/13 Kurohashi-Kawahara Lab., Kyoto University) URL: http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JECBasicSentenceData, 2013-02-01

[11] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html, 2013-02-01

corpus. Better criteria to measure similarity between clusters and new generated sentences should be explored in the future.

To summarize, the work presented in this paper demonstrates that it is possible to go beyond the old idea of constructing parallel corpora, and that it is possible to construct quasi-parallel corpora that lead to improvements in translation quality.

Future work may focus on finding a robust method to measure the similarity between analogical clusters and break sentences into phrases to apply the analogical technique to smaller pieces. We also intend to directly extract similar analogical clusters, instead of deriving them from the initial parallel corpus.

ACKNOWLEDGMENT

This work was partially supported by Foreign Joint Project funds from Kitakyushu Foundation for the Advancement of Industry, Science and Technology (FAIS).

REFERENCES

[1] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions (ACL 2007)*. Association for Computational Linguistics, 2007, pp. 177–180.

[2] D. S. Munteanu, A. Fraser, and D. Marcu, "Improved machine translation performance via parallel sentence extraction from comparable corpora." in *Proceedings of Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL 2004)*, 2004, pp. 265–272.

[3] D. S. Munteanu and D. Marcu, "Improving machine translation performance by exploiting non-parallel corpora," *Computational Linguistics*, vol. 31, no. 4, pp. 477–504, 2005.

[4] Y. Lü, J. Huang, and Q. Liu, "Improving statistical machine translation performance by training data selection and optimization." in *Proceedings of the 2007 Joint Meeting of the Conference on Empirical Methods on Natural Language Processing and the Conference on Natural Language Learning (EMNLP-CoNLL'07)*, 2007, pp. 343–350.

[5] Y. Marton, C. Callison-Burch, and P. Resnik, "Improved statistical machine translation using monolingually-derived paraphrases," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009): Volume-1*. Association for Computational Linguistics, 2009, pp. 381–390.

[6] A. Max, "Example-based paraphrasing for improved phrase-based statistical machine translation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*. Association for Computational Linguistics, 2010, pp. 656–666.

[7] Y. Lepage and E. 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 (IWP 05)*, 2005, pp. 57–64.

[8] P. Langlais, "Mapping source to target strings without alignment by analogical learning: A case study with transliteration," in *The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, 2013.

[9] W. Yang, H. Wang, and Y. Lepage, "Automatic acquisition of rewriting models for the generation of quasi-parallel corpus," in *Proceedings of the 6th Language and Technology Conference (LTC'13)*, 2013, pp. 409–413.

[10] A. Fujita, P. Isabelle, and R. 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)*. Association for Computational Linguistics, 2012, pp. 631–642.

[11] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant *et al.*, "Quantitative analysis of culture using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.

[12] M. Mangeot and E. Giguet, "Multilingual aligned corpora from movie subtitles," *Rapport technique, LISTIC*, 2005.

[13] Y. 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*. Association for Computational Linguistics, 1998, pp. 728–734.

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

[15] G. 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)*. Morgan Kaufmann Publishers Inc., 2002, pp. 138–145.

[16] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[17] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[18] S. Vogel, H. Ney, and C. Tillmann, "Hmm-based word alignment in statistical translation," in *Proceedings of the 16th conference on Computational Linguistics (COLING 96)-Volume 2*. Association for Computational Linguistics, 1996, pp. 836–841.

[19] A. Stolcke *et al.*, "SRILM-an extensible language modeling toolkit." in *INTERSPEECH*, 2002.

[20] P. Koehn, "Statistical significance tests for machine translation evaluation." in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, 2004, pp. 388–395.