Unsupervised and Semi-supervised Myanmar Word Segmentation Approaches for Statistical Machine Translation

Ye Kyaw Thu[†‡] **Andrew Finch**[†] **Eiichiro Sumita**[†] **Yoshinori Sagisaka**[‡]

[†]Multilingual Translation Lab.,
Universal Communication Research Institute,
National Institute of Information and Communications Technology, Kyoto, Japan

[‡]Global Information and Telecommunication Institute,
Department of Applied Mathematics Language and Speech Science Research Lab.,
Waseda University, Tokyo, Japan

{yekyawthu, andrew.finch, eiichiro.sumita}@nict.go.jp ysagisaka@gmail.com

Abstract

In statistical machine translation (SMT), word segmentation is generally a necessary step for languages that do not naturally delimit words. For many low-resource languages there are no word segmentation tools, and research on word segmentation for these languages is often quite scarce. In this paper, we study several plausible methods for Myanmar word segmentation for machine translation in order to shed light on promising avenues for future research. We propose a novel Bayesian learning method that can perform semi- supervised word segmentation with reference to a dictionary. We applied our method to the task of translating with Myanmar language, and compare our method to the following approaches to segmentation: human lexical/phrasal segmentation, character segmentation, syllable segmentation, purely unsupervised word segmentation, and the method of maximum matching. We found that unsupervised segmentation was the most effective segmentation. It outperformed maximum matching, which in turn was better than syllable segmentation.

Keywords: word segmentation, statistical machine translation, low resource languages, Myanmar language

1 Introduction

In many languages (for example Japanese, Chinese, and Myanmar which will be the focus of this paper), words are not necessarily delimited by white space in running text. For many natural language processing applications (for example machine translation), it is often useful or necessary to have text segmented into sequences of words. However, for many languages (Myanmar being one) there are no word segmentation tools, and there are two common approaches to dealing with this issue. The first is to apply unsupervised word segmentation tools to a body of monolingual text in order to induce a segmentation. The second is to use a dictionary of words in the language together with a set of heuristics to identify word boundaries in text.

Myanmar language can be accurately segmented into a sequence of syllables using rule based techniques. However, words composed of single or multiple syllables are not usually separated by white space. Although spaces are sometimes used for separating phrases for easier reading, it is not strictly necessary, and these spaces are rarely used in short sentences. There are no clear rules for using spaces in Myanmar language, and thus spaces may (or may not) be inserted between words, phrases, and even between a root word and its affixes. Myanmar language is a resource-poor language and large corpora, lexical resources, and grammatical dictionaries are not yet widely available. For this reason, using corpus-based machine learning techniques in developing a word segmentation tool is a challenging task.

As will be discussed in the following sec-

tion, one segmentation approach that has been successfully applied to Myanmar is that of training a machine translation from a sequence of syllables in Myanmar. Since syllables can be segmented accurately, this process has the advantage of having few segmentation errors while at the same time being one level above character- level segmentation. However, syllables are not equivalent to meaningful word units in the Myanmar language it is possible that if we can make word segmentation or near the word level, the quality of machine translation will increase.

2 Related Work

In this section, we will briefly introduce some proposed word segmentation methods with an emphasis on the schemes that have been applied to Myanmar.

Many word segmentation methods have been proposed especially for the Chinese and Japanese languages. These methods can be roughly classified into dictionary-based or rule-based and statistical methods [1], [2], [3], [4], [5]. In dictionary-based methods, only words that are stored in the dictionary can be identified and the performance depends to a large degree upon the coverage of the dictionary. New words appear constantly and thus, increasing size of the dictionary is a not a solution to the out of vocabulary word (OOV) problem. On the other hand, although statistical approaches can identify unknown words by utilizing probabilistic or cost-based scoring mechanisms, they also suffer from some drawbacks. The main issues are: they require large amounts of data; the processing time required; and the difficulty in incorporating linguistic knowledge effectively into the segmentation process [6]. For low-resource languages such as Myanmar, there is no freely available corpus and dictionary based or rule based methods are being used as a temporary solution.

If we only focus on Myanmar language word segmentation, as far as the authors are aware there have been only two published methodologies, and one study. Both of the proposed methodologies operate according using a process of syllable breaking followed by a process of Maximum Matching in which segmentation is performed using a dictionary and ambiguity is resolved by choosing the longest match. For complete explaination of Maximum Matching the reader is refered to [7]. The differences in the approaches come from the manner in which the segmentation boundary decision is made. In [8] statistical information is used (based on bigram information), whereas [9] utilize a word list extracted from a monolingual Myanmar corpus.

In a related study [10], various Myanmar word segmentation approaches including character segmentation, syllable segmentation, human lexical/phrasal segmentation, unsupervised and semi-supervised word segmentation, were investigated. They reported that the highest quality machine translation was attained either without word segmentation using simply sequences of syllables, or by a process of Maximum Matching with a monolingual dictionary. In this study the effectiveness of approaches unsupervised word segmentation using latticelm (with 3-gram to 7-gram language models) and supervised word segmentation using KyTea was evaluated, however, none of the approaches was able to match the performance of the simpler syllable/Maximum Matching techniques.

3 Methodology

3.1 Overview

The method we propose in this paper integrates dictionary-based word segmentation (similar to the maximum matching approaches used successfully in [8; 9; 10]) into a fully unsupervised Bayesian word segmentation scheme. Our idea is to combine the strengths of each of the two underlying approaches while at same time mitigating their weaknesses. Although the focus of our research is on the Myanmar language, we see no reason why a similar approach could not be applied to other low-resource languages in need of word segmentation tools.

Dictionary-based word segmentation has the advantage of being able to exploit human knowledge about the sequences of characters in the language that are used to form words. This approach is simple and has proven to be a very effective technique in previous studies. Problems arise due to the coverage of the dictionary. The dictionary may not be

able to cover the running text well, for example in the case of low-resource languages the dictionary might be small, or in the case of named entities, even though a comprehensive dictionary of common words may exist, it is likely to fall far short of covering all of the words that can occur in the language.

Unsupervised word segmentation techniques, have high coverage. They are able to learn how to segment by discovering patterns in the text that recur. The weakness of these approaches is that they have no explicit knowledge of how words are formed in the language, and the sequences they discover from text may simply be sequences in text that frequently occur and may bear no relationship to actual words in the language. As such these units, although they are useful in the context of the generative model used to discover them, may not be appropriate for use in an application that might benefit from these segments being words in the language. We believe that machine translation is one such application.

The proposed method aimed to give the unsupervised method a means of exploiting a dictionary of words in its training process. Our idea was to allow the integrated method to use the dictionary to segment text when appropriate, and otherwise use its unsupervised models to handle the segmentation. We do this by integrating a separate dictionary generation process into the generative model of the unsupervised segmenter to create a semi-supervised segmenter that segments using a single unified generative model. In the next section we briefly describe the Bayesian non-parametric segmentation component, and in the following section we explain how this was augmented with the dictionary segmentation model.

3.2 Bayesian Non-parametric Word Segmentation

The Bayesian non-parametric segmentation framework we used in our experiments a close relative to the approach proposed in [11].

Intuitively, the model has two basic components: a model for generating an outcome that has already been generated at least once before, and a second model that assigns a probability to an outcome that has not yet been produced. Ideally, to encourage the reuse of model parameters, the probability of generating a novel segment should be considerably lower than the probability of generating a previously observed segment. This is a characteristic of the Dirichlet process model we use and furthermore, the model has a preference to generate new segments early on in the process, but is much less likely to do so later on. In this way, as the cache becomes more and more reliable and complete, so the model prefers to use it rather than generate novel segments. The probability distribution over these segments (including an infinite number of unseen segments) can be learned directly from unlabeled data by Bayesian inference of the hidden segmentation of the corpus.

The underlying stochastic process for the generation of a corpus composed of segments $\mathbf{s}_k$ is usually written in the following from:

$$\begin{aligned} G|_{\alpha,G_0} &\sim DP(\alpha, G_0) \\ \mathbf{s}_k|G &\sim G \end{aligned} \tag{1}$$

G is a discrete probability distribution over the all segments according to a *Dirichlet process prior* with *base measure* G_0 and concentration parameter α. The concentration parameter $\alpha > 0$ controls the variance of G; intuitively, the larger α is, the more similar G_0 will be to G.

3.2.1 The Base Measure

For the base measure G_0 that controls the generation of novel sequence-pairs, we use a spelling model that assigns probability to new segments according to the following distribution:

$$\begin{aligned} G_0(\mathbf{s}) &= p(|\mathbf{s}|)p(\mathbf{s}|\,|\mathbf{s}|) \\ &= \frac{\lambda^{|\mathbf{s}|}}{|\mathbf{s}|!}e^{-\lambda V^{-|\mathbf{s}|}} \end{aligned} \tag{2}$$

where $|\mathbf{s}|$ is the number of tokens in the segment; V and is the token set size; and λ is the expected length of the segments.

According to this model, the segment length is chosen from a Poisson distribution, and then the elements of the segment itself is generated given the length. Note that this model is able to assign a probability to arbitrary sequences of tokens drawn from V.

3.2.2 The Generative Model

The generative model is given in Equation 3 below. The equation assignes a probability to the k^{th} bilingual segment $\mathbf{s}_k$ in a derivation of the corpus, given all of the other segments in the history so far $\mathbf{s}_{-k}$. Here $-k$ is read as: "up to but not including k".

$$p(\mathbf{s}_k|\mathbf{s}_{-k}) = \frac{N(\mathbf{s}_k) + \alpha G_0(\mathbf{s}_k)}{N + \alpha} \qquad (3)$$

In this equation, N is the total number of segments generated so far, $N(\mathbf{s}_k)$ is the number of times the segment $\mathbf{s}_k$ has occurred in the history. G_0 and α are the base measure and concentration parameter as before.

3.2.3 Bayesian Inference

We used a blocked version of a Gibbs sampler for training. In [12] they report issues with mixing in the sampler that were overcome using annealing. In [4] this issue was overcome by using a blocked sampler together with a dynamic programming approach. Our algorithm is an extension of application the forward filtering backward sampling (FFBS) algorithm [13] to the problem of word segmentation presented in [4]. We extend their approach to handle the joint segmentation and alignment of character sequences. We refer the reader to [4] for a complete description of the FFBS process. In essence the process uses a forward variable at each node in the segmentation graph to store the probability of reaching the node from the source node of the graph. These forward variables are calculated efficiently in a single forward pass through the graph, from source node to sink node (forward filtering). During backward sampling, a single path through the segmentation graph is sampled in accordance with its probability. This sampling process uses the forward variables calculated in the forward filtering step.

In each iteration of the training process, each entry in the training corpus was sampled without replacement; its segmentation was removed and the models were updated to reflect this. Then a new segmentation for the sequence was chosen using the FFBS process, and the models were updated with the counts from this new segmentation.

3.3 Incorporating the Dictionary

The dictionary was incorporated into the generative model for segmenting the corpus as an additional generative step. In this step the method of generating the segment was selected from either (1) generating the segment using the dictionary model, or (2) generating the segment using the Dirichlet process model described in the previous sections. The model selection process is governed by a probability p_{dict}, the prior probability that the segment will be generated from the dictionary. The dictionary model itself is a probability distribution over the words in the dictionary. The Gibbs sampling process descibed in Section 3.2.3 was extended to encompass this process. In other words, the generative story for our integrated model is as follows:

1. Generate all possible segments in all possible ways (both using the dictionary and the Dirichlet process model);
2. Choose one derivation from this set in accordance with its probability.

As the training process proceeds, the dictionary model induces a clustering over the segments in the corpus: those generated using the dictionary model and those generated using the Dirichlet process model. At the end of each iteration of the training the probability p_{dict} and the probability distribution over words in the dictionary were re-estimated from the counts based on the clusters to which the segments were assigned, and the occurrence counts of dictionary words in those portions of the corpus that were segmented using the dictionary model. We used add-one smoothing on the distribution over the dictionary words to ensure that their probabilities were never zero even if they did not occur in the segmentation of the corpus (since if their probability is zero, they cannot be sampled).

3.3.1 Initialization and Hyperparameter Inference

The sampling process is started from an initial random segmentation of the corpus. To obtain this segmentation, we set $p_{dict} = 0.5$ and set the dictionary model to be uniform distribution over the words in the dictionary. Then we ran a process of forward

filtering using only the base measure to assign probabilities to the segments, and sampled the initial segmentations with a backward sample step based on these probabilities.

In pilot experiments, the model did not appear to be particularly sensitive to the values of the λ parameter in the base measure. We set this value at 2 for all our experiments. The concentration parameter α was learned by sampling its value. Following [14] we used a vague gamma prior $Gamma(10^{-4}, 10^{4})$, and sample new values from a log-normal distribution whose mean was the value of the parameter, and variance was 0.3. We used the Metropolis-Hastings algorithm to determine whether this new sample would be accepted.

4 Experiments

For the experiments, we used three approaches: Bayesian learning without a dictionary, Bayesian learning with plain dictionary (here, plain means no prefix, suffix words or particles of Myanmar language like 'ed', 'er' in English) and Bayesian learning with a dictionary that included suffix words. We compared our proposed method to a partial segmentation provided by the original human translator, character segmentation, and syllable segmentation methods. For evaluation we used the BLEU score [15] rather than attempting to evaluate the segmentation performance directly due to: the lack of annotated reference data, the ambiguity of Myanmar segmentation, and our main objective was to improve the quality of Myanmar machine translation.

First, we segmented the Myanmar corpus using one of the segmentation methods. Second we divided the segmented corpus into training, development and test data for SMT processes. We used seven segmentation methods in total: "human translator's segmentation", "character segmentation","syllable segmentation", "Maximum Matching with plain dictionary", "Bayesian learning and segmentation without dictionary", "Bayesian learning and segmentation with plain dictionary", and "Bayesian learning and segmentation with plain dictionary plus common suffix Myanmar words". Character segmentation is segmentation into single characters, syllable segmentation is into sequenes of characters that form syllables; both are unambiguous.

We used a Myanmar-English Dictionary (26,421 words) and for getting common suffix Myanmar words we used an in-house corpus (161,993 sentences) built with news and blog data[16]. Standard Phrase Based Statistical Machine Translation (PBSMT) systems were trained using GIZA++ [17] for alignment, language modeling was done using IRSTLM version 5.80.01 [18]. Minimum error rate training (MERT) was used to tune the decoder's parameters and the decoding is done using the PBSMT system MOSES version 0.91 [19][20].

For evaluation, the output of all the systems was re-segmented into sequences of syllables, and then evaluated with reference to a syllable segmented reference set to allow the results to be cross-comparable.

4.1 Corpora

We used nine languages (Japanese, Korean, Indonesian, Malaysian, English, Chinese, French, Thai and Myanmar) from the multilingual Basic Travel Expression Corpus (BTEC), which is a collection of travel-related expressions [21]. We used already existing segmentation tools for all languages except Myanmar. For Myanmar language, we used the segmentation schemes described earlier. The source and target language corpus statistics are shown in Tables 1 and 2 respectively.

4.2 Results and Discussion

It can be seen from the results shown in Table 3 that the unsupervised segmentation has given rise to the highest BLEU scores. The performance of this approach greatly exceeds the performance of the previous published approaches on Myanmar segmentation: syllable segmentation and maximum matching. Clearly character-level segmentation is not effective for machine translation, and the human provided segmentation is also of little use on its own. Furthermore, the performance of the semi-supervised Bayesian segmenter exceeded the performance of the maximum matching approach in all experiments with the exception of Korean to Myanmar. Adding the suffix dictionary entries

Table 1. Corpus size in words for source languages

Source Language	Training	Development	Test
Japanese (ja)	594,127	95,727	85,337
Korean (ko)	559,243	89,519	80,819
Indonesian (id)	474,542	77,446	54,667
Malaysian (ms)	479,054	77,777	55,164
English (en)	527,268	86,934	75,002
Chinese (zh)	485,151	77,101	68,469
French (fr)	533,791	85,107	80,418
Thai (th)	512,054	86,401	77,602

Table 2. Corpus size in words for Myanmar language

Segmentation	Training	Development	Test
Human Segmentation	151,829	24,273	21,321
Character Segmentation	2,301,184	339,545	321,444
Syllable Segmentation	835,030	123,961	118,636
Maximum Matching with Plain Dictionary	711,776	102,347	97,554
Bayesian Segmentation without Dictionary	331,610	47,688	55,878
Bayesian Segmentation with Plain Dictionary	633,870	93,003	117,201
Bayesian Segmentation with Plain + Suffix Dictionary	606,630	86,639	81,244

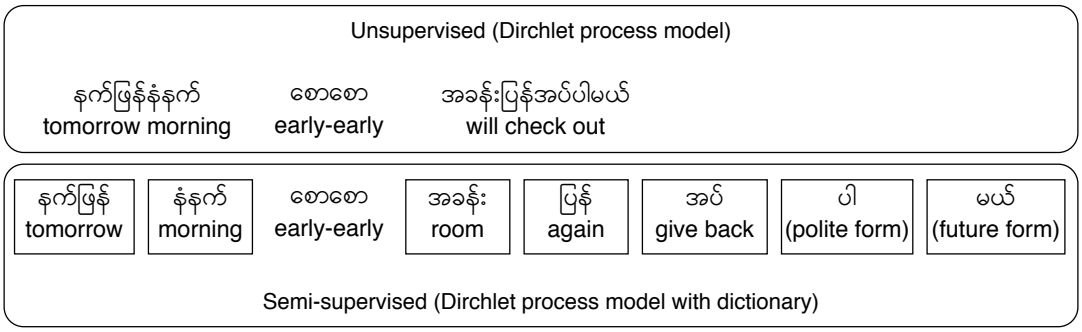

Figure 1. Unsupervised and supervised segmentation of a Myanmar sentence.

to the semi-supervised segmenter seemed to slightly degrade the performance in our experiments.

Visual inspection of the segmentation output from both unsupervised and semi-supervised processes showed that the semi-supervised process rarely made segmentation errors, and also that the dictionary model was used most of the time during the segmentation process. Figure 1 shows a typical example of a single sentence segmented by each method. The boxed segments in the semi-supervised segmentation indicate segments that were generated from the dictionary. The first difference is that "tomorrow morning" has been segmented as a single unit by the unsupervised segmenter rather than two individual words. This may be a useful unit for translation as it occurs reasonably frequently in the corpus and will simplify the alignment. "Early-early" is composed from the repetition of two syllables that are not in the dictionary, nor is the compound form. Both segmenters have identified this as a single unit. The biggest difference however is the segmentation of the expression "will check out", the unsupervised model has segmented it as a single word; the semi-supervised model has segmented as a sequence of words from the dictionary which although valid is probably not as useful for translation. The compound work meaning "checkout" was not contained in the dictionary.

Table 3. BLEU scores for human translator, character, syllable, Maximum Matching with plain dictionary, Bayesian without dictionary, Bayesian with plain dictionary, Bayesian with plain + suffix dictionary segmentations (bold numbers indicate the highest BLEU score)

Language Pair	Human	Character	Syllable	Maximum Matching	Bayesian No dict.	Bayesian Plain Dict.	Bayesian Suff. Dict.
ja-my	25.89	14.50	34.98	34.28	**39.00**	36.90	36.04
ko-my	27.77	14.21	33.64	35.53	**40.24**	35.18	34.64
id-my	13.02	3.85	13.83	14.00	**18.14**	15.93	16.89
ms-my	13.18	4.37	14.49	15.05	**19.64**	16.76	18.02
en-my	30.99	11.03	30.63	20.40	**37.26**	34.33	23.41
zh-my	27.26	11.54	31.40	31.82	**38.92**	32.45	32.31
fr-my	21.46	10.24	30.32	27.16	**33.14**	28.07	28.50
th-my	21.57	10.38	25.33	24.64	**28.65**	26.37	24.81

5 Conclusion

In this paper we studied the effectiveness of several approaches to word segmentation for Myanmar. In addition we contribute a novel word segmentation technique that marries unsupervised Bayesian non-parametric word segmentation with a dictionary-based approach within a single, unified model. This paper also contributes the first published evaluation of the quality of automatic translation from Indonesian, Malaysian, Chinese, French and Thai language to Myanmar language.

Previous approaches to Myanmar segmentation have shown promising results from syllable segmentation and from dictionary matching. This paper proposed an extension to approaches based on dictionary matching that allows the dictionary matching to occur within the framework of an unsupervised segmenter. While this approach proved more effective than a simple dictionary matching method, for Myanmar at least, a purely unsupervised approach proved to be the most effective strategy. We believe that this was due to a combination of the nature of the Myanmar language itself and the lack of many larger compound words in the dictionary.

We believe that improvements in Myanmar translation gained over the dictionary matching technique indicate that our approach may be effective for other unsegmented languages and intend to investigate this in future research.

References

[1] Zinmin Wu and Gwyneth Tseng. Chinese text segmentation for text retrieval: Achievements and problems. *ournal of the American Society for Information Science*, 44(5):532–542, 1993.

[2] Sun Maosong, Shen Dayang, and Benjamin K. Tsou. Chinese word segmentation without using lexicon and hand-crafted training data. In *Proceedings of the 17th international conference on Computational linguistics - Volume 2*, COLING '98, pages 1265–1271, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

[3] Papageorgiou and Constantine P. Japanese word segmentation by hidden markov model. In *Proceedings of the workshop on Human Language Technology*, HLT '94, pages 283–288, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[4] Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, pages 100–108, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[5] Chang Jyun-Shen, Chi-Dah Chen, and Shun-Der Chen. Chinese word segmen-

tation through constraint satisfaction and statistical optimization. In *Proceedings of ROC Computational Linguistics Conference*, pages 147–165, 1991.

[6] W. J. Teahan, Rodger McNab, Yingying Wen, and Ian H. Witten. A compression-based algorithm for chinese word segmentation. *Comput. Linguist.*, 26(3):375–393, September 2000.

[7] Wong Pak-kwong and Chan Chorkin. Chinese word segmentation based on maximum matching and word binding force. *Comput. Linguist.*, 1:200–203, 1996.

[8] Tun Thura Thet, Jin-Cheon Na, and Wunna Ko Ko. Word segmentation for the myanmar language. *J. Information Science*, 34(5):688–704, 2008.

[9] Hla Hla Htay and Kavi Narayana Murthy. Myanmar word segmentation using syllable level longest matching. In *IJCNLP*, pages 41–48, 2008.

[10] Ye Kyaw Thu, Andrew Finch, Yoshinori Sagisaka, and Eiichiro Sumita. A study of myanmar word segmentation schemes for statistical machine translation. *Proceeding of the 11th International Conference on Computer Applications*, pages 167–179, 2013.

[11] Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54, 2009.

[12] Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. Contextual dependencies in unsupervised word segmentation. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 673–680, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[13] Steven L Scott. Bayesian methods for hidden markov models : Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97(457):337–351, 2002.

[14] Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. A gibbs sampler for phrasal synchronous grammar induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*, pages 782–790, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[15] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. *Bleu: a Method for Automatic Evaluation of Machine Translation.* IBM Research Report rc22176 (w0109022), Thomas J. Watson Research Center, 2001.

[16] Ministry of Education. Department of the Myanmar Language Commission. 1993.

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

[18] Marcello Federico, Fondazione Bruno, Kessler Irst, and Mauro Cettolo. Efficient handling of n-gram language models for statistical machine translation.

[19] Franz J. Och. Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*, 2003.

[20] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowa, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007): demo and poster sessions*, pages 177–180, Prague, Czeck Republic, June 2007.

[21] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. Creating corpora for speech-to-speech translation. In *Proceedings of EUROSPEECH-03*, pages 381–384, 2003.