Integrating Dictionaries into an Unsupervised Model for Myanmar Word Segmentation

Ye Kyaw Thu
NICT
Keihanna Science City
Kyoto, Japan
yekyawthu@nict.go.jp

Andrew Finch
NICT
Keihanna Science City
Kyoto, Japan
andrew.finch@nict.go.jp

Eiichiro Sumita
NICT
Keihanna Science City
Kyoto, Japan
eiichiro.sumita@nict.go.jp

Yoshinori Sagisaka
GITI/Speech Science Research Lab.
Waseda Univerity
Tokyo, Japan
ysagisaka@gmail.com

Abstract

This paper addresses the problem of word segmentation for low resource languages, with the main focus being on Myanmar language. In our proposed method, we focus on exploiting limited amounts of dictionary resource, in an attempt to improve the segmentation quality of an unsupervised word segmenter. Three models are proposed. In the first, a set of dictionaries (separate dictionaries for different classes of words) are directly introduced into the generative model. In the second, a language model was built from the dictionaries, and the n-gram model was inserted into the generative model. This model was expected to model words that did not occur in the training data. The third model was a combination of the previous two models. We evaluated our approach on a corpus of manually annotated data. Our results show that the proposed methods are able to improve over a fully unsupervised baseline system. The best of our systems improved the F-score from 0.48 to 0.66. In addition to segmenting the data, one proposed method is also able to partially label the segmented corpus with POS tags. We found that these labels were approximately 66% accurate.

1 Introduction

In many natural language processing applications, for example machine translation, parsing and tagging, it is essential to have text that is segmented into sequences of tokens (these tokens usually represent 'words'). In many languages, including the Myanmar language (alternatively called the Burmese language), Japanese, and Chinese, words are not necessarily delimited by white space in running text. However, in some low-resource languages (Myanmar being one) broad-coverage word segmentation tools are scarce, 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 finite state automata (examples being (Berment, 2004; Thu et al., 2013a)). 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 to develop word segmentation tools is a challenging task.

Proceedings of the 5th Workshop on South and Southeast Asian NLP, 25th International Conference on Computational Linguistics, pages 20–27,
Dublin, Ireland, August 23-29 2014.

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 Thai, Khmer, Lao, Chinese and Japanese languages. These methods can be roughly classified into dictionary-based (Sornlertlamvanich, 1993; Srithirath and Seresangtakul, 2013) and statistical methods (Wu and Tseng, 1993; Maosong et al., 1998; Papageorgiou and P., 1994; Mochihashi et al., 2009; Jyun-Shen et al., 1991). 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 (Teahan et al., 2000). 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 Maximum Matching; the differences in the approaches come from the manner in which the segmentation boundary decision is made. In (Thet et al., 2008) statistical information is used (based on bigram information), whereas (Htay and Murthy, 2008) utilize a word list extracted from a monolingual Myanmar corpus.

In a related study (Thu et al., 2013a), 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.

In (Pei et al., 2013) an unsupervised Bayesian word segmentation scheme was augmented by using a dictionary of words. These words were obtained from segmenting the data using another unsupervised word segmenter. The probability distribution over these words was calculated from occurrence counts, and this distribution was interpolated into the base measure.

3 Methodology

3.1 Baseline Non-parametric Bayesian Segmentation Model

The baseline system, and the model that forms the basis for all of the models is a non-parametric Bayesian unsupervised word segmenter similar to that proposed in (Goldwater et al., 2009). The major differences being the sampling strategy and the base measure. The principles behind this segmenter are described below.

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 re-use of model parameters, the probability of generating a novel segment should be considerably lower then 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.

3.1.1 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 the set of tokens V (in this paper V is the set of all Myanmar syllables). The motivation for using a base measure of this form, is to overcome issues with overfitting when training the model; other base measures are possible for example the enhancement proposed in Section 3.4.

3.1.2 The Generative Model

The generative model is given in Equation 3 below. The equation assignes a probability to the k^{th} 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”.

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

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.

3.1.3 Bayesian Inference

We used a blocked version of a Gibbs sampler for training. In (Goldwater et al., 2006) they report issues with mixing in the sampler that were overcome using annealing. In (Mochihashi et al., 2009) 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 (Scott, 2002) to the problem of word segmentation presented in (Mochihashi et al., 2009). We extend their approach to handle the joint segmentation and alignment of character sequences. We refer the reader to (Mochihashi et al., 2009) 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. The two hyperparameters, the Dirichlet concentration parameter α, and the Poisson rate parameter λ were set by slice sampling using vague priors (a Gamma prior in the case of α and the Jeffreys prior was used for λ). The token set size V used in the base measure was set to the number of types in the training corpus, and $V = 3363$.

3.2 Dictionary Augmented Model

The dictionary augmented model is in essence the same model as proposed by (Thu et al., 2013b), but a different dictionary was used. Their method integrates dictionary-based word segmentation (similar to the maximum matching approaches used successfully in (Thet et al., 2008; Htay and Murthy, 2008; Thu et al., 2013a)) into a fully unsupervised Bayesian word segmentation scheme.

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.

This method gives the unsupervised method a means of exploiting a dictionary of words in its training process, by allowing the integrated method to use the dictionary to segment text when appropriate, and otherwise use its unsupervised models to handle the segmentation. To do this a separate dictionary generation process is integrated into the generative model of the unsupervised segmenter to create a semi-supervised segmenter that segments using a single unified generative model.

3.3 Dictionary Set Augmented Model

In this model, the a set of subsets of the dictionary were extracted based on the part-of-speech labels contained in the dictionary (Lwin, 1993). This set of subsets was not a partition of the original dictionary since some of the types in the dictionary were ambiguous causing some overlap of the subsets. In the previous model, during the generative process a decision was made, with a certain probability learned from the data, as to whether the segment would be generated from the unsupervised sub-model or the dictionary sub-model. In this model, the decision to generate from the dictionary model is refined into a number of decisions to generate from a number of subsets of the dictionary, each with its own probability. These probabilities were re-estimated from the sampled segmentation of the corpus at the end of each iteration of the training (in a similar manner to the dictionary augmented model). A diagram showing the generative process is shown in Figure 1.

3.4 Language Model Augmented Model

In (Theeramunkong and Usanavasin, 2001) dictionary approaches were deliberately avoided in order to address issues with unknown words. Instead a decision tree model for segmentation was proposed. Our approach although different in character (since a generative model is used), shares the insight that knowledge of how words are constructed is key to segmentation when dictionary information is absent.

In this model we used the dictionary resource, but in a more indirect manner. We use a language model to capture the notion of exactly what constitutes a segment. To do this words in the dictionary were first segmented into syllables. Then, a language model was trained on this segmented dictionary. This model will assign high probabilities to the words it has been trained on, and therefore in some sense is able to capture the spirit of the dictionary-based methods described previously. However, it will also have learned something about the way Myanmar words are composed from syllables, and can be expected to assign a higher probability to unknown words that resemble the words it has been trained on, than to sequences of syllables that are not consistent with its training data.

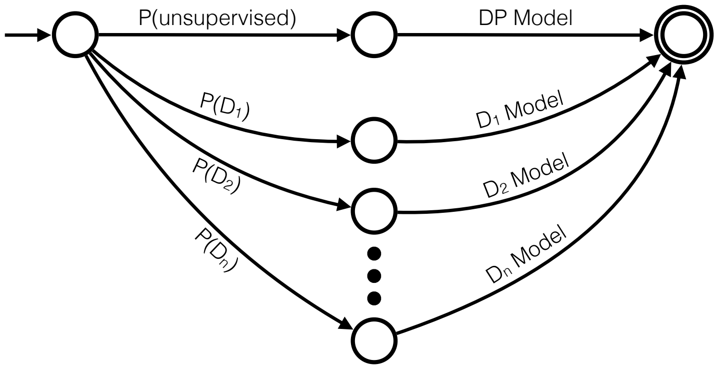

Figure 1: Generative process with multiple dictionaries competing to generate the data alongside an unsupervised Dirichlet process model.

This model can be naturally introduced directly into the Dirichlet process model as a component of the base measure. Equation 2 decomposes into two terms:

1. A Poisson probability mass function: $\frac{\lambda^{|\mathbf{s}|}}{|\mathbf{s}|!}e^{-\lambda}$
2. A uniform distribution over the vocabulary: $\frac{1}{|V|^{|\mathbf{s}|}}$.

The first term above models the choice of length for the segment. The second term models the choice of syllables that comprise the segment is in essence a unigram language model that provides little information about segments are constructed, and serves simply to discourage the formation of long segments that would lead to overfitting. We directly replace the part of the base measure with our more informative language model built from the dictionary.

4 Experiments

4.1 Overview

In the experimental section we aim to analyze two aspects of the performance of the proposed segmentation approaches. Firstly their segmentation quality and secondly for those approaches that are capable of partially labeling the corpus, the accuracy of the labeling.

4.2 Corpora

For all our experiments we used a 160K-sentence subset of the Basic Travel Expression (BTEC) corpus (Kikui et al., 2003), for the Myanmar language. This corpus was segmented using an accurate rule-based approach into individual syllables, and this segmentation was used as the base segmentation in all our experiments.

In addition a test corpus was made by sampling 150-sentences randomly from the corpus. This corpus was then segmented by hand and the segments were annotated manually using the set of POS tags that our system was capable of annotating, together with an ‘UNK’ tag to annotate segments that fell out of this scope. The test sentences were included in the data used for the training of the Bayesian models. These sentences were treated in the same manner the rest of the data in that they were initially syllable segmented. At the end of the training, the test sentences together with the segmentation assigned by the training, were extracted from the corpus, and their segmentation/labeling was evaluated with reference to the human annotation.

4.3 Segmentation Performance

We used the Edit Distance of the Word Separator (EDWS) to evaluate the segmentation performance of the models. This technique yields precision, recall and F-score statistics from the edit operations required to transform one segmented sequence into another by means of the insertion, deletion and substitution of segmentation boundaries. In this measure the substitution operator corresponds to an identity substitution and therefore indicates a correct segment boundary. Insertions correspond to segmentation boundaries in

Method	Precision	Recall	F-score	Sub	Ins	Del
Unsupervised	82.27	33.82	0.48	348	681	75
Maximum Matching	78.39	99.42	0.88	1023	6	282
Dictionary	89.67	57.34	0.70	590	439	68
Dictionary Set	89.46	51.99	0.66	535	494	63
Language Model (LM)	88.15	31.10	0.46	320	709	43
Dictionary Set + LM	91.27	49.76	0.64	512	517	49

Table 1: Segmentation Performance.

the segmented output that do not correspond to any segmentation boundary in the reference. Deletions correspond to segmentation boundaries in the reference that do not correspond to any segmentation boundary in the segmented output. Precision recall and F-score are calculated as follows using the Chinese Word Segmentation Evaluation Tookit (Joy, 2004):

$$\text{precision} = \frac{\text{\#substitutions}}{\text{\#segment boundaries in output}}$$

$$\text{recall} = \frac{\text{\#substitutions}}{\text{\#segment boundaries in reference}}$$

$$\text{F-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

Table 1 shows the segmentation performance all of the systems. In terms of precision we see an improvement when the additional models are added to the baseline unsupervised model. The maximum matching strategy has the lowest precision but the highest recall, this is due to the over-generation of segmentation boundaries in regions where no dictionary matches are possible. In these regions the reference segmentation boundaries are always annotated (since the model defaults to syllable segmentation in these regions), but at the expense of precision. This is reflected in the relatively low numbers of insertions (6), and the relatively high number of deletions (282). As expected the dictionary set approach gave similar performance to the Dictionary approach. The language model approach produced a respectable level of precision, but a low value for recall. When integrated into the dictionary-based models however, it was able to increase precision.

4.4 Labeling Accuracy

We evaluated the accuracy of the labeling on two methods: the first method used only a set of dictionaries (as described in Section 3.3). This method was able to label with an accuracy of 64.53%. The second method consisted of the same technique, but with the addition of the language model trained on the syllable segmented dictionaries (as described in Section 3.4). We found that the dictionary-based language model was able to improved the labeling accuracy 1.2% to 65.73%.

5 Examples and Analysis

Figure 2 shows an example an unsupervised segmentation with a typical error. Frequent words are often attached to neighboring words to form erroneous compound segments. In this example, taken from real output of the segmenter, the word ‘De’ (this) has been attached to the word ‘SarOuk’ (book), and similar the words in the phrase ‘KaBeMharLe’ (where is it) have all been segmented as a single segment (which occurs frequently in the corpus).

In Figure 3, a typical segmentation from the maximum matcher is shown. In this example the word ‘BarThar’ (language) occurs in the dictionary but the word ‘JaPan’ (Japan) does not. The maximum matcher defaults to segmenting the word for Japan into its component syllables, whereas the unsupervised segmenter with dictionary has attempted an unsupervised segmentation on this part of the string. The word ‘Japan’ occurs sufficiently frequently in the BTEC corpus that the segmenter has been able to

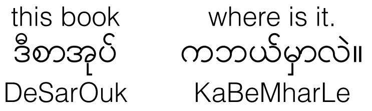

Figure 2: An unsupervised segmentation.

learn the word during training and has thereby managed to successfully segment the word in the output. The word for language was segmented by means of the embedded dictionary model.

Figure 4 shows an example of the partial labeling produced from the model that used a set of dictionaries in combination with a dictionary-based language model in the base measure. Due to the small amount of resources available, substantial parts of the sequence are unable to be labeled (and are annotated with the 'U' tag in the figure, indicating that they were segmented by the unsupervised component of the model). The remainder of the words are annotated with POS tags corresponding to the dictionary they were generated from.

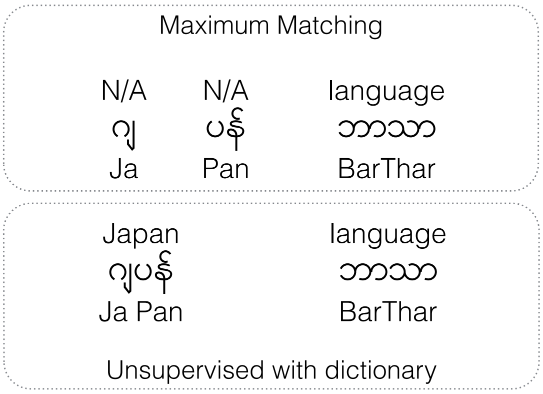

Figure 3: A segmentation from maximum matching.

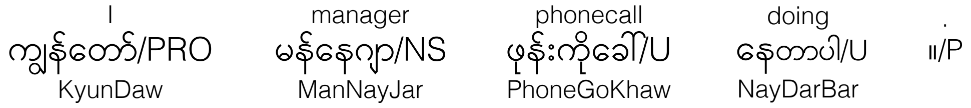

Figure 4: A partially labeled segmentation.

6 Conclusion

In this paper we have proposed and investigated the effectiveness of several methods intended to exploit limited quantities of dictionary resources available for low resource languages. Our results show that by integrating a dictionary directly into an unsupervised word segmenter we were able to improve both precision and recall. We found that attempting to model word formation using a language model on its own was ineffective compared with the approaches that directly used a dictionary. However, this language model proved useful when used in conjunction with the direct dictionary-based models, where it served to assist the modeling of words that were not in the dictionary. In future work we intend to develop the dictionary set approach by extending it to introduce basic knowledge of the morphological structure of the language directly into the model.

References

Vincent Berment. 2004. Sylla and gmsword: applications to myanmar languages computerization. In *Burma Studies Conference*.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2006. 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. Association for Computational Linguistics.

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

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

Joy, 2004. *Chinese Word Segmentation Evaluation Toolkit*.

Chang Jyun-Shen, Chi-Dah Chen, and Shun-Der Chen. 1991. Chinese word segmentation through constraint satisfaction and statistical optimization. In *Proceedings of ROC Computational Linguistics Conference*, pages 147–165.

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

San Lwin. 1993. *Myanmar - English Dictionary*. Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar.

Sun Maosong, Shen Dayang, and Benjamin K. Tsou. 1998. 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. Association for Computational Linguistics.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. 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. Association for Computational Linguistics.

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

Wenzhe Pei, Dongxu Han, and Baobao Chang. 2013. A refined hdp-based model for unsupervised chinese word segmentation. In Maosong Sun, Min Zhang, Dekang Lin, and Haifeng Wang, editors, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, volume 8202 of *Lecture Notes in Computer Science*, pages 44–51. Springer Berlin Heidelberg.

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

Virach Sornlertlamvanich. 1993. Word segmentation for thai in machine translation system. *Machine Translation, National Electronics and Computer Technology Center, Bangkok*, pages 50–56.

A Srithirath and P. Seresangtakul. 2013. A hybrid approach to lao word segmentation using longest syllable level matching with named entities recognition. In *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2013 10th International Conference on*, pages 1–5, May.

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

Thanaruk Theeramunkong and Sasiporn Usanavasin. 2001. Non-dictionary-based thai word segmentation using decision trees. In *In Proceedings of the First International Conference on Human Language Technology Research*.

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

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

Ye Kyaw Thu, Andrew Finch, Eiichiro Sumita, and Yoshinori Sagisaka. 2013b. Unsupervised and semi-supervised myanmar word segmentation approaches for statistical machine translation.

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