IMPROVING WORD SEGMENTATION FOR THAI SPEECH TRANSLATION

Paisarn Charoenpornsawat and Tanja Schultz

InterACT, Language Technologies Institute
Carnegie Mellon University
{paisarn,tanja}@cs.cmu.edu

ABSTRACT

A vocabulary list and language model are primary components in a speech translation system. Generating both from plain text is a straightforward task for English. However, it is quite challenging for Chinese, Japanese, or Thai which provide no word segmentation, i.e. the text has no word boundary delimiter. For Thai word segmentation, Maximal Matching, a lexicon-based approach, is one of the popular methods. Nevertheless this method heavily relies on the coverage of the lexicon. When text contains an unknown word, this method usually produces a wrong boundary. When extracting words from this segmented text, some words will not be retrieved because of wrong segmentation. In this paper, we propose statistical techniques to tackle this problem. Based on different word segmentation methods we develop various speech translation systems and show that the proposed method can significantly improve the translation accuracy by about 6.42% BLEU points compared to the baseline system.

Index Terms **—** Spoken language translation, Speech Recognition, Text Processing, Word Segmentation.

1. INTRODUCTION

Unlike English and most other Western languages, Thai text provides no word boundary delimiters. Thus, before performing any natural language processing tasks, Thai requires a preprocessing step which breaks a sequence of characters into words. Several word segmentation approaches have been proposed for Thai such as lexicon-based approaches, rule-based approaches, statistical approaches and machine-learning-based approaches [1, 2, 3, 4]. Besides lexicon-based approaches, other approaches require either linguistic knowledge or manually segmented text. Therefore, Maximal Matching (MM) [3], a lexicon-based approach, is widely applied to several Thai speech and language applications [5, 6].

Although, MM does not need any linguistic knowledge or any segmented text, it heavily relies on the coverage of the lexicon. It mostly produces proper segmentation for general text when all the words are covered by the lexicon. However, for unknown words, i.e. new words which are not in the language (most often these are names, foreign words or loanwords), this technique usually generates wrong word boundary. It could break an unknown word into many tokens composing of words and non-words.

In developing our Thai speech translation system, we have to segment training data to generate a vocabulary list and a language model for an Automatic Speech Recognition (ASR) system and a Statistical Machine Translation (SMT) system. Thus if we are applying MM for the segmentation, some words, which are not covered by the segmenter's lexicon, will be separated into several (often non-word) tokens. This has an impact on the generation of pronunciation of these words and, in addition has the effect that these words will not be included in the pronunciation dictionary of the ASR system. Consequently, these words cannot be recognized by the system and thus increase the out-of-vocabulary (OOV) rate of the ASR system and errors in the speech translation.

To alleviate the problem of MM, we apply statistical techniques which can be learnt from unsegmented text and do not require any linguistic knowledge or segmented text. From our experience, pure statistical approaches are outperformed by MM when the words are covered by the lexicon. Therefore we integrate statistical techniques into MM to overcome both its limitations. Though, we are focusing on an impact of word segmentation on speech translation, it is also better to consider whether the word segmentation has any effect on text translation. Thus in our experiments, we will show the impact on speech recognition, speech translation, and on text translation.

2. PROBLEMS OF UNKNOWN WORD BOUNDARIES

Thai unknown words can be formed by a combination of known words and unknown strings (non-words). For example, ไมโครซอฟต์ (Microsoft) is composed of ไม, โค (ox), ร, ซอ (fiddle), and ฟต์. Only the 2nd and 4th tokens are known words. The others are non-words which have no meaning. Also Thai unknown words can be composed of one or more known words such as บุญเสริม (a person name) is composed of บุญ (merit) and เสริม (to strengthen).

For unknown words consisting of two or more words are very challenging in detecting and mostly happens in Thai names. Since Thai data in our experiments are translated from English, it rarely contains Thai proper names. However, there are a lot of foreign words and loan words which usually contain an unknown string. Thus, we focus in this paper on unknown words that have unknown strings.

3. THAI WORD SEGMENTATION ALGORITHMS

3.1. Maximal Matching (MM)

MM will generate all possible segmentations from a given sentence based on a provided lexicon and then selects the best segmentation which has the fewest number of words. For example, there are 2 possible segmentations for "*Iwanttobeafireman*":

I.) *I want to be a fire man* (7 words),
II.) *I want to be a fireman* (6 words).

In this example, "*I want to be a fireman*" will be selected because it has fewer words than the other. When there is an unknown word in a sentence (i.e. not covered by the lexicon), MM might not segment it correctly. For example, suppose "*Los Angeles*" in an unknown word. The sentence "*IwanttogotoLosAngeles*" will be segmented as "*I want to go to Los Angel es*". Because there is no "*Los Angeles*" but there is a word "*angel*" in the lexicon, MM will split "*LosAngeles*" into "*Los Angel es*". Both "*Los*" and "*es*" are unknown strings.

When MM produces an unknown string (i.e. the string is not covered by the lexicon), it indicates the existence and location of an unknown word. We then can employ this information to apply other techniques to identify a boundary of the unknown word.

3.2. Left-to-Right Entropy

Entropy information has been successfully applied to Thai syllable segmentation [7] and word extraction [8]. This technique does not require any language knowledge neither a lexicon. In this paper, we propose a new word segmentation technique using entropy information. First, we define left and right conditional entropy information as shown respectively in Equation 1 and 2.

$$LE(c_{i,j}) = -\sum_{\forall x \in A} p(x \mid c_{i,j}) * \log_2 p(x \mid c_{i,j}) \quad (1)$$

$$RE(c_{i,j}) = -\sum_{\forall z \in A} p(z \mid c_{i,j}) * \log_2 p(z \mid c_{i,j}) \quad (2)$$

The variable "$c_{i,j}$" is a substring starting from the character at i and ending at j. The variables "x" and "z" are characters at position i-1 and j+1, respectively. The variable "A" is the set of all possible characters in the script of the corresponding language.

Intuitively, the entropy of a word is higher than the entropy of any substrings inside the word. Starting from the same point when increasing the size of a substring, the right entropy will be reduced except when it reaches the end of a word. If we start from right to left, when increasing the size of the substring, the left entropy of the substring will be decreasing except when it reaches the beginning of a word. From this intuition, we include both the left and right entropy in the calculation [8].

Applying the left and right conditional entropy to find a word boundary is straightforward. Starting from left to right, find the substring $c_{1,n}$ which satisfies $RE(c_{1,n}) > RE(c_{1,n-1})$. $c_{1,n}$ could be consider as a word. However, sometimes this is not true. It is better also to check the left entropy of the next word. From our experience, instead of considering only $c_{1,n}$ as a word, it is better to consider $c_{1,n+1}$ as well. For example, suppose a string "$c_1c_2c_3c_4c_5c_6c_7c_8$" is separated into two words "$c_1c_2c_3c_4$" and "$c_5c_6c_7c_8$". If $RE(c_{1,1}) > RE(c_{1,2})$ and $RE(c_{1,2}) < RE(c_{1,3})$, we will consider $c_{1,3}$ as a word based on the right entropy. If this is correct, c_4 should be the starting point of the new word. So $LE(c_{4,8})$ should be higher than $LE(c_{5,8})$. Suppose in this case $LE(c_{5,8})$ is much higher than $LE(c_{4,8})$. Then, we make a decision by including both left and right entropy in calculation. In the case, we compare the values between $RE(c_{1,3})+LE(c_{4,8})$ and $RE(c_{1,4})+LE(c_{5,8})$ and select the one that has the highest value. In this case $RE(c_{1,4})+LE(c_{5,8})$ should have a higher value.

3.3. Left-to-Right Entropy with Mutual Information

Applying only the left-to-right entropy technique is still not sufficient because it sometimes breaks a word into many tokens. For example, by using this technique, the word "อินโดนีเซีย" (Indonesia) is segmented into two words, "อินโด" (Indo) and "นีเซีย" (nesia), although the two substrings belong to one word. However, this problem could be relieved by using Mutual Information (MI) as defined in Equation 3.

$$MI(x, y) = \log_2 \frac{p(xy)}{p(x) * p(y)} \quad (3)$$

MI has been successfully applied to identify whether two instances x and y are belong to the same word or not. If so, we will combine them into one word.

From the previous example, MI of "อินโด" (Indo) and "นีเซีย" (nesia) is given as 2.68, which is considerably high, based on our data. In our experiments, two strings will be dependent if MI is equal or higher than 1.35 which is manually derived from the unlabeled training data.

3.4. Maximum Average Entropy per Word

Instead of generating a word boundary when the right entropy considerably increases as described in section 3.1,

we try to find the best segment which can provide the highest average of both left and right entropy per word. Because this technique actually finds the overall highest entropy for every word instead of using local entropy, it usually outperforms the left-to-right entropy technique. The problem could be defined as in the equation 4.

$$W_i = \arg\max_{W_i} \frac{\sum_{j=1}^{n} \left(RE(w_j) + LE(w_j) \right)}{n} \quad (4)$$

In Equation 4, $W_i = w_1 w_2 .. w_n$ is a sequence of words and and w_j is a sequence of characters while $RE(w_j)$ and $LE(w_j)$ are the same as defined in equation 1 and 2.

4. EXPERIMENTS

4.1. Data

Four experiments we used about 300,000 bilingual Thai-English aligned sentences from the medical and tourism domain including BTEC [9]. For Thai, we applied the open-source word segmentation program called SWATH [10] using the described MM technique. Since the accuracy of MM depends on its lexicon, we replaced the default lexicon in SWATH with the lexicons from both Lex*i*tron and RI [11] to have better coverage which gives 2.5% in the OOV rate. To test our approaches, we randomly selected 500 bilingual sentences for a test set where every Thai sentence has at least one unknown string. An unknown string was generated by MM, when there was an unknown word. Thus every sentence in the test set has at least one unknown word. After that we asked a Thai native speaker to read and record the 500 Thai sentences and used them as the test data for ASR systems.

4.2. Thai Speech Recognition

For every ASR system, we use the same acoustic models trained from about 90 hours of recording from about 150 speakers. The acoustic component is based on context-dependent models using quintphone (±2) with 2000 acoustic models and 32 Gaussians per codebook. However, the developed systems differ in the pronunciation dictionary and the language model according to the different word segmentation techniques. We applied the described word segmentation techniques to the training data and then constructed a word list and a language model. From the word list, we applied the example-based grapheme-to-phoneme conversions for Thai [7] to generate a pronunciation dictionary. In the experiments, we built 5 ASR systems as follows:

1.) *Baseline system*: used a segmented text generated by MM to create the word list and to train the language model,

2.) *Oracle system*: applied the same technique as in the baseline system to segment text, however we manually added unknown words composed of all unknown strings from the test set to the segmenter's lexicon,

3.) *Left-to-right entropy (L2R-ENT) system*: employed segmented text from the baseline system but resegmented an area having an unknown string with the left-to-right entropy technique,

4.) L*eft-to-right entropy + MI (L2R-ENT+MI) system*: same as the left-to-right entropy system but instead of using only the left-to-right entropy technique, we also integrated MI to combine words together which was already described in 3.4,

5.) *Maximum average entropy per word (MA-ENT) system*: same as the left-to-right entropy system but instead using the maximum average entropy per word technique.

SYSTEM	WER	DEL	OOV	VOC
Baseline	29.00%	1.92%	8.88%	19.9K
Oracle	13.74%	2.18%	0.35%	20.4K
L2R-ENT	22.07%	2.48%	5.14%	24.8K
L2R-ENT + MI	20.52%	2.91%	4.60%	26.0K
MA-ENT	20.94%	3.77%	3.14%	26.9K

Table 1: the results for different ASR systems

Table 1 shows the results of the described five ASR systems where DEL and VOC are a deletion error rate and a vocabulary size, respectively. The results show that different techniques solve the unknown word problem in different levels. L2R-ENT technique can reduce the OOV rate from 8.88% to 5.14% and leads to absolute reduction of 6.03% in WER compared to the baseline system. Applying MI on top of L2R-ENT decreases the OOV rate further to 4.60% and reduces the WER by another 1.55% absolute compared to the L2R-ENT system. MA-ENT provides the lowest OOV rate but the system generates slightly higher WER than L2R-ENT+MI system.

This is because MA-ENT sometimes combines more than one word together. It usually combines function words together or a function word with a content word. For example, for the word "สเต็ก" (steak) in some context the boundary is produced correctly, while sometimes MA-ENT prefers to combine it with another word to get higher entropy such as the word "สเต็กกับ" which combines "steak" and "with". Thus both "steak" and "steakwith" will be included in the pronunciation dictionary. Therefore, when recognizing the utterance "I want steak with …", sometimes it recognizes "steakwith" as one word, thus producing a deletion error. This can explain why the WER of MA-ENT is higher than L2R-ENT system.

4.3. Thai Statistical Machine Translation

To build a Thai SMT system, we used the CMU statistical machine translation toolkit [12]. The system is based on phrase-to-phrase extraction. Phrase extraction is done using Pharaoh [13]. The CMU decoder retrieves all possible word

and phrase translations from the given input to generate the translation lattice and then searches for the best translation which gives the high probability.

In our experiments, we built 5 different systems based on the word segmentation techniques corresponding to the five ASR systems as described above. These systems were trained on the 300,000 bilingual sentences and tested on 500 sentences with 1 reference translation.

Table 2 shows the BLEU scores of the different SMT systems for two different tasks: 1.) Speech-to-Text (S2T) translation and 2.) Text-to-Text (T2T) translation. S2T translation was using the ASR system output as source sentences. T2T translation was using the text reference as the input. Thus the S2T results are lower than the T2T results because the ASR system propagated errors to the translation system.

SYSTEM	S2T (%)	T2T (%)
Baseline	34.00	46.21
Oracle	41.89	47.76
L2R-ENT	39.08	46.58
L2R-ENT + MI	40.02	43.64
MA-ENT	40.43	45.63

Table 2: The results from different SMT systems in BLEU

In the S2T translation task, the improvement should be more than 4.0% in BLEU for 95% confidence level. Our proposed techniques can significantly improve the BLEU score by about 5-6% absolute compared to the baseline system. The results of S2T systems show the same trend as the results of ASR systems except for the L2R-ENT+MI and MA-ENT systems. Even though MA-ENT gives slightly higher WER in ASR task, it still gives a better performance in the S2T task. This is because SMT can generate proper translation even MA-ENT combines two or more words together.

In T2T translation task, the word segmentation is not a severe problem because SMT can handle wrong segmentations. Even though, an unknown word in the segmentation is split into several tokens, SMT sill can produce proper translation. It is encouraging to see that the translation performance of the MA-ENT system comes very close to the optimal results of the Oracle system. In other words, our MA-ENT technique almost eliminates the lack of lexicon coverage in Thai word segmentation for the purpose of speech translation.

5. CONCLUSIONS

In this paper we presented three different techniques to solve the unknown word problem in the lexicon-based word segmentation approach. We constructed five ASR and SMT systems based on different segmentation techniques. From the experimental results, the MA-ENT word segmentation technique produced the lowest OOV rate for the ASR system. Even though MA-ENT system had slightly higher WER than L2R-ENT+MI system, MA-ENT system still provided the best result in S2T task. In T2T task, the unknown word problem in the segmentation, however, is not a problem.

Although, we achieved significant performance gains from MA-ENT technique, there is still room for improvement. Besides the unknown word problem, there is an ambiguity problem in the segmentation which could affect both pronunciation and meaning. In our future work, we will investigate this ambiguity problem by considering both pronunciation and meaning together. Additionally we can apply this technique to other Asian languages such as Chinese, Lao or Khmer.

6. REFERENCES

[1] W. Aroonmanakun, "Collocation and Thai Word Segmentation," In Proc. of the 5th SNLP&COCOSDA Workshop. Pathumthani, Thailand, 2002.

[2] T. Theeramunkong and S. Usanavasin, "Non-Dictionary-Based Thai Word Segmentation Using Decision Trees," In Proc. of HLT. San Diego, USA. 2001.

[3] S. Meknavin, P. Charoenpornsawat and B. Kijsirikul, "Feature-based Thai word segmentation," In Proc. of NLPRS, 1997.

[4] C. Haruechaiyasak, S. Kongyoung, and M. N. Dailey. A Comparative Study on Thai Word Segmentation Approaches. In Processdings of ECTI-CON. 2008.

[5] S. Suebvisai and et.al., "Thai Automatic Speech Recogntion," In Proc. of ICASSP, Philadelphia, PA. 2005.

[6] S. Dangsaart et.al., "Intelligent Thai text - Thai sign translation for language learning," Comput. Educ. 51, 3, Nov. 2008.

[7] P. Charoenpornsawat and T. Schultz, "Example-Based Grapheme-to-Phoneme Conversion for Thai," In Proc. of Interspeech, Pittsburgh PA. 2006.

[8] V. Sornlertlamvanich, T. Potipiti, and T. Charoenporn, "Automatic corpus-based Thai word extraction with the c4.5 learning algorithm," In Proc. of the 18th Conference on Computational Linguistics - Volume 2. 2000.

[9] T. Takezawa et.al., "Towards a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," In Proc. of LREC, Canary Islands, Spain, 2002.

[10] P. Charoenpornsawat, SWATH: Smart Word Analysis for Thai. http://www.cs.cmu.edu/~paisarn/software.htm. 2003.

[11] Lexitron Version 2.0. Thai-English Dictionary. Source available: http://lexitron.nectec.or.th. 2003.

[12] S. Vogel et.al., "The CMU statistical Machine Translation System" In Proceedings of MT summit IX, 2003.

[13] P. Koehn. "Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models," In Proc. of AMTA, Washington, DC. 2004.