Advanced MT Seminar Report
Word Segmentations and their integrations within MT

Thuy Linh Nguyen (thuylinh@andrew.cmu.edu)

May 6, 2008

1 Introduction

In machine translation, the smallest unit of a model is a word and a sentence is a sequence of words separated by spaces. Unlike English and other western languages, Asian language such as Chinese, Japanese, Thai, etc..., there are no space boundaries between words. This posts a problem when translating these languages to English. Due to the similarity of the problem, researchers have focused on Chinese as the major language of the word segmentation problem resarch. The solutions to Chinese language is also applicable to other languages. In this report, we will review the existing researches on word segmentations for machine translation. There are three approaches to the problem of word segmentation for machine translation.

In the conventional approach, a word segmentation program segments a sequence of characters in to words and then machine translation models apply to the segmented data. In this way, the issues shift to a single language segmentation problem.

The second approach exploits character translation instead of word segmentation. A single language segmentation requires an external word segmentation toolkit together with a dictionary of Chinese word. Xu et al. [2004] investigated the translation at the character level and minimized the resource needed for building a translation system by learning the word list from Chinese characters alignment with English words. The word segmentation used the learned word list instead of an external dictionary.

The third approach combines word segmentation to the decoder into one package. The word segmentation toolkit might contain errors and do not optimize with respect to the translation task. Xu [2005] represented a test sentence as lattices of different segmentations instead of a single best segmentation and integrated the lattices into the machine translation's search for the best translation.

In this report we will review the existing three directions with focus on the last two. The next section will discuss the single language word segmentation research, the section 3 will review character translation. Section 4 will review

the integration of word segmentation into lattice translation. The last two sections, section 5 and 6, will be our conclusion and future direction.

2 Word Segmention

2.1 Word Segmentation problems

First of all, there is no widely accepted definition of word boundary in Chinese. Sproat et al. [1994] conducted an experiment including six people individually segmented the same Chinese text. The consistency of their results is only 75%, the number might be even lower if more people involved into the task. For word segmentation research, we put aside the controversy of word boundary definition but rather use a pre-segmented corpus as the reference, the word segmentation task still need to tackle the problems of ambiguities and words never seen before.

2.1.1 Ambiguity

One of the main difficulty in Chinese word segmentation is the lack of unambiguous word boundary indicator in the text. In Figure 1, the first character of the first example is also a word by itself. In the second example, the same character is a middle character of a compound word. In the last example, the character is the rightmost character of a word.

Position	Example
Left	产生 'to come up with'
Word by itself	产 小麦 'to grow wheat'
Middle	生产线 'assembly line'
Right	生产 'to produce'

Figure 1: A Chinese character (hanzi) can occur in different positions (Xue [2003])

2.1.2 Unknown words

In conventional approach, word segmentation models often use a dictionary as a reference to detect the word boundary. However a word list can not list all the possible words in the vocabulary. There are several ways to introduce new words in English. For example, a combination of existing words creates a new word. Also, personal names are created by combination of characters in unpredictable manner. The transliteration of foreign name added an other source of unknown words in Chinese.

A word segmentation method should be able to resolve the ambiguity and detect words that are not in the training data.

2.2 Word segmentation methods

Various methods have been proposed to address the problem of word segmentation in previous study. They can be classified in to three categories:

- Purely dictionary-based approach
- Purely statistical-based approach
- Statistical based approach using manual segmentation data.

The main topic of this paper is word segmentation for machine translation therefore we will only review the most prominent researches in each approach.

2.2.1 Purely dictionary based

Cheng et al. [1999] proposed maximum matching heuristics that detect the word boundary based on a given word list. The algorithm greedily search the sentence from the left to right for the longest sequence of characters that match a word in the dictionary and insert the word boundary to that point. The greedy search starts again that new point of word boundary. The heuristic is very simple to implement and it can resolve the ambiguity in many cases. But its performance is completely dependent on the coverage of the dictionary and unable to detect unknown words in the corpus.

2.2.2 Purely Statistical-based

The statistical-based word segmentation relies on the mutual information of adjacent characters to detect the boundaries of words (Sun et al. [1998]). A group of characters that have mutual informations greater than a threshold form a word. Statistical based approach does not relies on a dictionary and it can train on any unsegmented data. Peng and Schuurmans [2001] used unsupervised machine learning model to learn the probability of a word in the dictionary. The dictionary therefore extracted from the data. The dictionary was further revised by removing words that have low mutual informations between characters of the words. The advantage of purely statistical approach is that it does not require an external dictionary and can apply to any languages. The drawback of is that the purely statistical approach does not perform as well in term of segmentation accuracy.

2.2.3 Statistical based approach using manual segmentation data

The statistical approach using manual segmentation data got many attentions in word segmentation research. Xue [2003] tags Chinese characters in to one of four tags and applied maximum entropy to learn the character tagging model. Zhang et al. [2003] incorporated Chinese word segmentation, part-of-speech tagging, disambiguation and unknown words recognition into a whole theoretical frame.

The state-of-the-art word segmentation that the current machine translation uses as the preprocessing tool using Conditional Random Field(CRF). This approach was first presented in Peng et al. [2004] and later extended in the open source Stanford word segmenter Tseng et al. [2005].

They tagged the characters that begin a new word with tag START(S), the characters in the middle or end of a word with tag NONSTART(NS) as depicted in Figure 2. The task of segmenting a new sentence becomes the problem of assigning a sequence of tags to the sequence of Chinese characters.

Figure 2: Word Segmentation as Character Tagging

Let $\mathbf{c} = (c_1, c_2, \ldots, c_K)$ be a Chinese sentence, the CRF probability that $\mathbf{t} = (t_1, t_2, \ldots, t_K)$ be the character tags of $\mathbf{c}$ is

$$\Pr(\mathbf{t}|\mathbf{c}) = \frac{1}{\mathrm{Z}(\mathbf{c})} \exp\left(\sum_{k=1}^{k=K} \sum_{i} \lambda_i f_i(t_{k-1}, t_k, \mathbf{c}, k)\right)$$

The feature functions $f_i(t_{k-1}, t_k, \mathbf{c}, k)$ are state transition features, character lexical features.

Using character tag for word segmentation problem has the advantage of combining the ambiguity resolution and unknown word detection into one package.

Sighan Bakeoff 2003	*F-score Tseng et al (2005)*	*F-score Peng et al. (2004)*
CTB	0.863	0.849
AS	0.970	0.956
HK	0.947	0.928
PK	0.953	0.941

Figure 3: Word Segmentation Results

Figure 3 displays both segmenter's F-score on the closed track in Sighan bakeoff 2003 evaluation. The Stanford segmentation result is on the left column. The evaluation on four different test sets of different domain. The word segmentation accuracies are very high in all four test sets. Given the current still low quality of Chinese-English translation, we believe that even if the word segmentation toolkit could achieve higher accuracy, the better word segmentation tool kit is not guaranteed for a better translation quality.

3 Translation without word segmentation

Xu et al. [2004] presented a new method for Chinese-English translation without

using a manual segmented corpus on a predefined dictionary. They learned the word list dictionary from the Chinese characters alignment to English.

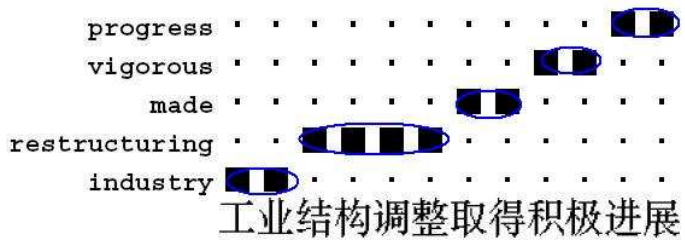

Figure 4: Grouping character into words from Character-English Alignment

First characters of Chinese corpus are separated by spaces and aligned with English parallel text. The output of the alignment is an alignment matrix for each sentence pair as in Figure 4. The contiguous Chinese characters that align to the same English word form a Chinese word. In the example, the first two characters aligned to "industry" form a word, the next four characters aligned to "restructuring" form the second word of the sentence,etc... This process automatically segmented the training corpus and also generated a dictionary of Chinese words.

They compared the three models: translation with no segmentation (where each character is a word), translation with segmentation from self learned dictionary and the translation with standard LDC segmentation. The bilingual training data was the Chinese treebank and its English translation. Table 1 displays all three system results when evaluated on the NIST 2003 evaluation.

Method	Error rates		Accuracy
	WER	PER	BLEU
No segment	73.3	56.5	27.6
Learned segment	70.4	54.6	29.1
LDC segment	71.9	54.4	29.2

Table 1: Translation performance of different segmentation methods

The translation result of learned segmentation is similar to the translation with LDC segmentation, and the translation without segmentation is worse. They concluded it is possible to achieve the same performance without using an external dictionary.

3.1 Discussion

The paper presented a method for Chinese word segmentation based on its character English alignment but its performance does not outperform the standard word segmentation. The entries in the dictionary are only words in the training data. Therefore, the word segmentation accuracy of a test sentence

decreases if there are many character sequences in the test sentence but not in the training sentence. A segmentation with the combination dictionary of LDC and the learned dictionary would be a possible extension experiment. Also, the LDC word segmentation is purely dictionary based with greedy longest match searching heuristics, the longest sequence of characters that match a word in the dictionary might not be the word segmented to align with an English word in the reference. To resolve this problem, the system should use the rich context features CRF segmentation model(e.g Stanford word segmentation) trained on the learned word segmented corpus to re-segment the training and testing data.

4 Integrated Chinese Word Segmentation in SMT

Figure 5: Test sentence as a sequence of characters

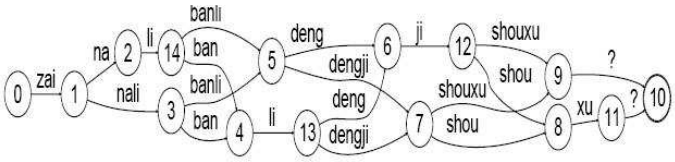

Figure 6: Segmentation lattice without weights

The single best segmentation for a test sentence might contain errors and not optimal for translation. Xu [2005] addressed this issue by translate all the segmentation alternatives of the test sentence instead of only the best segmentation. We will use the example in the paper to illustrate the idea.

Suppose we have a Chinese sentence written in pinyin format "zai na li ban li deng ji shou xu ?" Figure 5 depicts the sequence of characters of the sentence, the label of the arc from node $i-1$ to node i is the character i-th of the sentence. In lattice translation, the test sentence is converted to lattice format as in Figure 6, each arc of the lattice is a sub-sequence of characters that form a word in the dictionary, an path through the lattice is a possible word segmentation of the sentence.

One of the problem with the lattice translation is that when searched for the best translation path, the decoder biased toward short lattice paths. They composed the unweighted lattice with a language model weighted finite state transducer(FST). The weights of the language model FST trained on a pre-segmented Chinese corpus with the SRILM toolkit. The output of the composition is a weighted lattice as in Figure 7

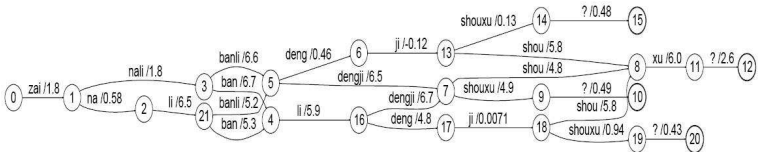

Figure 7: Segmentation lattice with weights

They reported the translation results on the IWSLT 03 test set. The training corpus consists of 20K Chinese-English sentence pairs. They used two translation systems: a finite state transducer translation with monotone decoding and a phrase-based translation system[1]. The finite state transducer system output very short translations because the model can not include the word penalty feature so its result is very low in compare with phrase-based system. We will discuss here the results of the phrase-based system only.

Seqmentation methods	WER[%]	PER[%]	NIST	BLEU
Single best segmentation	53.6	43.8	8.18	38.9
Segmentation lattice without weight	47.0	38.1	8.09	40.2
Segmentation lattice with weight	47.2	38.0	8.18	40.4

Table 2: Translation with different segmentation methods

Table 2 displays the phrase-based translation results for different segmentation methods, the lattice translation outperformed the single best segmentation. The model does not benefit from assigning weights to lattices.

4.1 Discussion

In this section we want to analyze why the lattice translation without weights already outperformed the single best translation. Note that in Chinese segmentation, the longest matching heuristics efficiently resolve the ambiguities in many case. This heuristics also tends to generate the short segmentation. Therefore, for Chinese translation, short segmentations do not cause unweighted lattice translation lower score in compare with single best segmentation system. An other way, the system does not benefit from weighted lattice translation.

When there are unseen words in the single segmentation, the translation system benefit from the lattice representation, it will search for the translation of short words instead of giving an unknown word combination of several characters. When the single best segmentation of the test sentence consists only words already in the training data, because of the longest match heuristics the translation decoder will bias toward the single best analysis. In this paper, the model was trained on small 20K training corpus with high number of unknown words. The lattice translation might not benefit when the training data is large and covers the test set.

[1]The paper did not report the reordering window of the phrase-based system

5 Conclusion and Future Direction

This report have reviewed current research how word segmentation could benefit machine translation. While the single language word segmentation is very success and could achieve very high accuracy, little research has focused on the integration word segmentation into machine translation framework. That explains why current state-of-the-art Chinese-English translation system still uses an external segmenter as preprocessing and applied the MT model afterward.

Two researches that attempting to incorporate part of word segmentation task into translation that we discussed above are experimented on a very small training corpus and these experiments not yet gave promising results. We believe that an extension of Xu et al. [2004] with a richer context of word segmentation model could have higher benefit to the translation model.

Another direction of the research is incorporating word alignment and segmentation into one framework so that the data has better alignment with English corpus. The weight estimation as described in Eisner [2002] could provide a principle approach to this possible future research.

References

K. S. Cheng, G. H. Young, and Wong. A study on word-based and integral-bit chinese text compression algorithms. *Journal of the American Society for Information Science*, 50(3):218–228, 1999.

Jason Eisner. Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–8, Philadelphia, July 2002. URL `http://cs.jhu.edu/ jason/papers/#acl02-fst`.

Fuchun Peng and Dale Schuurmans. Self-supervised Chinese word segmentation. *Lecture Notes in Computer Science*, 2189:238+, 2001. URL `citeseer.ist.psu.edu/peng01selfsupervised.html`.

Fuchun Peng, Fangfang Feng, and Andrew Mccallum. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of Coling 2004*, pages 562–568, Geneva, Switzerland, Aug FebruaryMarch–Aug FebruaryJuly 2004. COLING. URL `http://www.aclweb.org/anthology-new/C/C04/C04-1081.bib`.

Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. A stochastic finite-state word-segmentation algorithm for chinese. In *Meeting of the Association for Computational Linguistics*, pages 66–73, 1994. URL `#`.

M. Sun, D. Shen, and B. Tsou. Chinese word segmentation without using lexicon and hand-crafted training data. In *Proc. of COLING-ACL '98*, pages 1265–1271, 1998. URL `citeseer.ist.psu.edu/sun98chinese.html`.

Huihsin Tseng, Chang Pichuan, Andrew Galen, and Christopher Manning Daniel Jurafsky. A conditional random field word segmenter. 2005. URL http://www.aclweb.org/anthology-new/W/W06/W06-0100.pdf.

Xu. Integrated chinese word segmentation in statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 141–147, Pittsburgh, PA, October 2005.

J. Xu, R. Zens, and H. Ney. Do we need chinese word segmentation for statistical machine translation? In *Proceedings of the Third SIGHAN Workshop on Chinese Language Learning*, pages 122–128, Barcelona, Spain, July 2004.

N. Xue. Chinese word segmentation as character tagging. *International Journal of Computational Linguistics and Chinese Language Processing*, 2003. URL http://www.aclclp.org.tw/clclp/v8n1/v8n1a2.pdf.

Hua-Ping Zhang, Hua-Ping Zhang, Qun Liu, Xue-Qi Cheng, Hao Zhang, and Hong-Kui Yu. Chinese lexical analysis using hierarchical hidden markov model. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, pages 63–70, Morristown, NJ, USA, 2003. Association for Computational Linguistics.