Unsupervised Word Segmentation Without Dictionary

Jason S. Chang
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road,
Hsinchu, 300, Taiwan, ROC

jschang@cs.nthu.edu.tw

Tracy Lin
Department of Communication Engineering
National Chiao Tung University
1001, Ta Hsueh Road,
Hsinchu, 300, Taiwan, ROC

tracylin@faculity.nctu.edu.tw

This prototype system demonstrates a novel method of word segmentation based on corpus statistics. Since the central technique we used is unsupervised training based on a large corpus, we refer to this approach as *unsupervised word segmentation*.

The unsupervised approach is general in scope and can be applied to both Mandarin Chinese and Taiwanese. In this prototype, we illustrate its use in word segmentation of Taiwanese Bible written in Hanzi and Romanized characters. Basically, it involves:

- Computing mutual information, MI, between Hanzi and Romanized characters *A* and *B*. If *A* and *B* have a relatively high MI, we lean toward treating *AB* as a word.

- Using a greedy method to form words of 2 to 4 characters in the input sentences.

- Building an N-gram model from the results of first-round word segmentation

- Segmenting words based on the N-gram model

- Iterating between the above two steps: building N-gram and word segmentation

Computing mutual information. Using mutual information is motivated by the observation of previous work by Hank and Church (1990) and Sproat and Shih (1990). If *A* and *B* have a relatively high MI that is over a certain threshold, we prefer to identify *AB* as a word over those having lower MI values. In the experiment with Taiwanese Bible, the system identified Hanzi and Romanized syllables. Out of those, we obtained pairs of consecutive single or double Hanzi characters and Romanized syllables. So those pairs are commonly known as character bigrams, trigrams, and fourgrams. We differed from the common N-gram calculation and treated those as pairs of character sequence in order to apply mutual information statistics. Table 1 shows some examples of the pairs and MI values. We have excluded pairs having MI 2.2 or lower.

Table 1. Example of consecutive pairs (C_1, C_2) and MI values for the text of Taiwanese Bible

C_1	C_2	Mutual Information
婦	仁人	568.6012
婦	仁	281.1248
果	子	34.9152
婦仁	人	34.6398
園內	樹	23.6275
仁	人	16.4376
內	樹裡	10.7914
園	內	10.6569
通	食	8.9151
樹	裡	4.2192
仁人	對	3.2395
阮	通食	2.8967

Word Segmentation. With the potential words and MI values indicating their likelihood, we proceeded to segment the text of a large corpus into words. For the Taiwanese Bible, we had to take care of the problem of text being written down in more than one writing system: we had mixed Hanzi and Romanized syllables as input. Using a greedy method, we gradually formed words of 2 to 4 characters (or Romanized syllables) in the input sentences. A word with high-MI constituent characters took precedence in forming words.

Table 2. Example of consecutive pairs and MI values for the text of Taiwanese Bible

Left Syllable String, C_1	Right Syllable String, C_2	Mutual Information MI(C_1, C_2)
2-Syllable pairs		
婦	仁	281.1248
果	子	34.9152
仁	人	16.4376
園	內	10.6569
通	食	8.9151
樹	裡	4.2192
仁人	對	3.2395
3-Syllable pairs		
婦	仁人	568.6012
婦仁	人	34.6398
園內	樹	23.6275
內	樹裡	10.7914
阮	通食	2.8967

When successive words were formed, they could not contradict with the words determined previously. For instance, given the input "婦仁人對蛇講：「園內樹裡的果子阮通食," we looked

up the table storing MI statistics and obtained the information shown in Table2. First, we formed words of two characters. Based on the information in Table 2, the system formed the words, 婦仁, 果子, 通食, 園內, 樹裡. Notice that 仁人 is not selected because of confliction with previous decision about the word 婦仁. Subsequently, we tried to extend the two-syllable words chosen. A word is extended to three or four syllables if the MI is increased and in the corpus over τ % of instances the two-character words can be extended that way. Currently, we set $\tau = 60$.

Admittedly, there is limitation to what distributional regularity based on MI can be exploited for word segmentation and there were still many errors in the first-round word segmentation results. For instance, for the input, "我祈禱耶和華講：『主耶和華啊 … ," the system produced the segmentation of "我 / 祈禱 / 耶和華 / 講 / ： / 『 / 主耶 / 和華 / 啊 /." The first instance of 耶和華 was segmented correctly, while the second instance of 耶和華 was over-segmented because of the frequent character 主 before it. That problem can be partially alleviated by an Expectation Maximization Algorithm of learning an N-gram model of word segmentation.

Building an N-gram model.

Currently, we used the unigram model where the probability of each word was estimated based on the Good-Turing smoothing technique. First we tally the total number of words N and count R of each word W. Let N_r be the number of distinct words have count r. Also, let N_0 be the number of distinct syllable strings that never appear as a word. Godd-Turing smoothing stipulates that we calculate r' as an adjustment for r as follows:

$$r_0 = N_1 / N_0$$

$$r_i = (i+1) N_{i+1} / N_i$$

After the adjustment step, we obtained the probability for the unigram model as follows:

$$P(W) = r' / N \quad \text{where } r' \text{ is the smoothed count of } W$$

  For instance, we had the counts after the first-round MI-based segmentation as showed in Table 3.

Table 3. Good-Turing estimates for unigrams: Adjusted frequencies and probabilities

r	N_r	r*	$P_{GT}(\cdot)$

0	972,444	**0.00417**	0.0000000074
1	**4,056**	**0.97436**	0.0000017360
2	**1,976**	**1.37854**	0.0000024562
3	908	2.94273	0.0000052431
4	668	3.69760	0.0000065881

Table 4. Probabilities used in word segmentation of "主耶和華"

Word	Raw Count	Probability
主耶	**268**	0.0004775030
和華	**433**	0.0007714881
主	**1,048**	0.0018672506
耶和華	**5,612**	0.0099990557
主耶和	**0**	0.0000000074
華	404	0.0007198180
耶和	**1**	0.0000017360

Word Segmentation based on the N-gram model. We proceeded to redo the word segmentation task on the same corpus with an aim of rectifying the errors occurring in the previous stage. This was done following the standard dynamic programming procedure of Viterbi Algorithm of finding segmentation *S* satisfying the following optimality condition:

$$S = \underset{(W_1..W_n)}{\arg\max} \prod_{i=1,n} P(W_i).$$

For the example of "我祈禱耶和華講：『主耶和華啊 … " given earlier, the system is likely to produce correct segmentation "我 / 祈禱 / 耶和華 / 講 / ： / 『 / 主 / 耶和華 / 啊 /… ."

Table 5. Probabilities for various segmentations of "主耶和華"

Segmentation, *S*	P(*W₁*)	P(*W₂*)	P(*W₃*)	P(*S*)
主, 耶和華	0.0018672506	0.0099990557	-	0.0000186707
主耶, 和華	**0.0004775030**	**0.0007714881**	-	**0.0000003684**
主耶和, 華	**0.0000000074**	**0.0007198180**	-	**0.0000000000053**
主, 耶和, 華	**0.0018672506**	**0.0000017360**	**0.0007198180**	**0.0000000000023**

Iterating between building N-gram and word segmentation. The improved word segmentation obviously will bring about a better N-gram model for segmentation. Subsequently the improved N-gram will help to produce segmentation results of higher accuracy. The process of improvement usually converges quickly after a couple of iterations.

Our demonstration prototype sheds new lights on the extensively studied problem of word

segmentation. The prototype illustrates:

- It is possible to achieve high-precision word segmentation for a sufficiently large corpus without a dictionary, rivaling human annotation.

- The heuristic MI-based approach by Sproat can be extended effective to handle word longer than two characters

- A more theorectically sound approach based on N-gram model and unsupervised learning based on EM-like algorithm can bring about higher performance than the heuristic approach based on mutual information.

- Unsupervised, self-organized word segmentation can provide an objective view of word segmentation. This should be considered as a quantitative, corpus-dependent method when setting up a segmentation standard or benchmark for word segmentation.

High-precision segmentation of Hanzi text can be achieved by unsupervised training on a reasonably sized corpus. Unsupervised word segmentation represents an innovative way to acquire lexical units in a large corpus based on lexical distributional regularity. Word segmentation algorithm is standard Viterbi algorithm and is independent of N-gram trained on the corpus, making it easy to change domains. The approach is useful in an indefinite number of areas, and lends itself to customization for a particular user or task. For example, the results can be used to prepare a concordance, as the first steps in many natural language processing systems such as machine translation, information retrieval, or text-to-speech system. Finally, the model explored here can be a basis for self-organized word segmentation and alignment of bilingual Chinese-English corpus.

Acknowledgements

References
Church, K and P. Hank, "Word Association Norms, Mutual Information, and Lexicography," Computational Linguistics, 16:1, 1990, pp. 22-29.

Sproat, Chinese Word Segmentation, *First International Conference on Language Resources & Evaluation: Proceedings*, 1998, pp. 417— 420.

Richard Sproat, Chilin Shih, 1990, A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese & Oriental Languages, 4(4): 336-351.

R. Sproat, Chilin Shih, William Gale, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for chinese. Computational Linguistics, 22(3): 377–404.