Word Segmentation for Chinese Wikipedia Using N-Gram Mutual Information

Ling-Xiang Tang[1], *Shlomo Geva*[1], *Yue Xu*[1] *and Andrew Trotman*[2]

[1]School of Information Technology
Faculty of Science and Technology
Queensland University of Technology
Queensland 4000 Australia

{l4.tang, s.geva, yue.xu}@qut.edu.au

[2]Department of Computer Science
Universityof Otago
Dunedin 9054 New Zealand

andrew@cs.otago.ac.nz

Abstract *In this paper, we propose an unsupervised segmentation approach, named "n-gram mutual information", or NGMI, which is used to segment Chinese documents into n-character words or phrases, using language statistics drawn from the Chinese Wikipedia corpus. The approach alleviates the tremendous effort that is required in preparing and maintaining the manually segmented Chinese text for training purposes, and manually maintaining ever expanding lexicons. Previously, mutual information was used to achieve automated segmentation into 2-character words. The NGMI approach extends the approach to handle longer n-character words. Experiments with heterogeneous documents from the Chinese Wikipedia collection show good results.*

Keywords Chinese word segmentation, mutual information, n-gram mutual information, boundary confidence

1 Introduction

Modern Chinese has two forms of writings: simplified and traditional. For instance, the word China is written as 中国 in simplified Chinese, but as 中國 in traditional Chinese. Furthermore, a few variants of Chinese language exist in different locales including: Mainland China, Taiwan, Hong Kong, Macau, Singapore and Malaysia. For instance, a laser printer is called 激光打印机 in mainland China, but 鐳射打印機 in Hongkong, and 雷射印表機 in Taiwan.

In digital representations of Chinese text different encoding schemes have been adopted to represent the characters. However, most encoding schemes are incompatible with each other. To avoid the conflict of different encoding standards and to cater for people's linguistic preferences, Unicode is often used in collaborative work, for example in Wikipedia articles. With Unicode, Chinese articles can be composed by people from all the above Chinese-speaking areas in a collaborative way without encoding difficulties. As a result, these different forms of Chinese writings and variants may coexist within same pages. Besides this, Wikipedia also has a Chinese collection in Classical Chinese only, and versions for a few Chinese dialects. For example, 贛語(Gan) Wikipedia, 粵語(Cantonese) Wikipedia and others. Moreover, in this Internet age more and more new Chinese terms are coined at a faster than ever rate. Correspondingly, new Chinese Wikipedia pages will be created for the explanations of such terms. It is difficult to keep the dictionary up to date due to the rate of creation and extent of new terms. All these issues could lead to serious segmentation problems in Wikipedia text processing while attempting to recognise meaningful words in a Chinese article, as text will be broken down into single character words when the actual n-gram word can not be recognised. In order to extract n-gram words from a Wikipedia page, the following problems must be overcome:

- Mix of Chinese writing forms: simplified and traditional
- Mix of Chinese variants
- Mix of Classical Chinese and Modern Chinese
- Out of vocabulary words

Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.

There may be two options to tackle the these new issues in Chinese segmentation: (1) use existing methods and solutions; or (2) attempt new technique. In general, on the basis of the required human effort, the Chinese word segmentation approaches can be classified in two categories:

- Supervised methods, e.g. training-based, or rules-based methods, which require specific language knowledge. Normally, a pre-segmented corpus is employed to train the segmentation models e.g. PPM [13], or word lexicons need to be prepared for dictionary-based methods e.g. CRF [10].
- Unsupervised methods, which are less complicated, and commonly need only simple statistical data derived from known text to perform the segmentation. For instance, statistical methods using different mutual information formulas to extract two-character words rely on the bi-gram statistics from a corpus [2] [12].

The drawbacks of supervised methods are obvious. The effort of preparing the manually segmented corpus and parameters tuning is extensive. Also, the selected corpus mainly from modern Chinese text source may only cover a small portion of Wikipedia Chinese text. Plus, out-of-vocabulary(OOV) words are problematic for dictionary based methods. Different writing and different variant can lead to different combinations of characters representing the same word. Furthermore, according to the 2nd International Chinese Word Segmentation Bakeoff result summary [1], the rankings of participants results in different corpora not being very consistent which may indicate that the supervised methods used in their segmentation system are form (simplified or traditional) sensitive. To make use of these existing systems, the segmentation could be done the by converting all Chinese text into one unified form, simplified Chinese, for example. However, the resulting performance may be cast in doubt because the Chinese form conversion could not change the way the variant is used radically. For example, 鐳(Radium) character in 鐳射(laser, or 激光 simplified Chinese equivalent), will remain the same after such conversion, and 鐳射 would still not be recognised correctly as a word for the simplified-Chinese oriented segmentation system. At the time of writing, no performance of segmentation targeted on Chinese Wikipedia corpus using the-state-of-the-art systems are reported.

To avoid the effort of preparing and maintaining segmented text and lexicons for different corpora and potential issues when applying existing methods on Chinese Wikipedia articles, a simple unsupervised statistical method called n-gram mutual information(NGMI), which relies on the statistical data from text mining on Chinese Wikipedia corpus, is proposed in this paper. We extend the use of character-based mutual information to be segment-based in order to realize n-gram Chinese word segmentation. To achieve this goal, we introduce a new concept named boundary confidence(BC) which is used to determine the *boundary* between segments. The n-gram words are thus separated by the boundaries. The estimation of boundary confidence is based on the mutual information of adjoining segments. Since n-gram mutual information looks for boundaries in text but not for words directly, it overcomes the limitation of traditional usage of mutual information in the recognition of bi-gram words only.

2 Previous Studies

Pure statistical methods for word segmentation are less well studied in the Chinese segmentation research. They are the approaches that make use of statistical information extracted from text to identify words. The text itself is the only ``training'' corpus used by the segmentation models.

Generally, the statistical methods used in Chinese segmentation can be classified into the following groups: Information Theory (e.g. entropy and mutual information), Accessory Variety, t-score and Others. The accuracy of the segmentation is commonly evaluated using the simple recall and precision measures:

$$R = \frac{c}{N},\ and\ P = \frac{c}{n}$$

R is the recall rate of the segmentation
P is the precision rate of the segmentation
c is the number of correctly identified segmented words
N is the number of unique correct words in the test data
n is the number of segmented words in the test data

In a recent study, an accessory variety(AV) method has been proposed by Feng et al. [4] to segment words in a unsupervised manner. Accessory variety measures the probability of a character sequence being a word. A word is separated from the input text by judging the independence of a candidate word from the rest by using accessor variety criterion in considering the number of distinct preceding and trailing characters. An AV value of a candidate word is the minimal number of distinct preceding or trailing characters. The higher the number, the more independent the word is.

Information theory can help group character sequences into words. Lua [7] [8] and Gan [8] used the entropy measure in their word segmentation algorithm. A character sequence is a possible word if its overall entropy is lower than the total entropy of individual characters. Using this entropy theory for word judgment differently, Tung and Lee [14] considered the relationship of a candidate word with all possible preceding and trailing single-characters appearing in the corpus. The entropy values are calculated for those characters given that they occur in either the left hand side or the right hand side of this candidate word. If entropy values on either side are high, the candidate word could be an actual word. Mutual information and its derived algorithms are mainly used in finding bi-gram words.

Generally, mutual information is used to measure the strength of association for two adjoining characters. The stronger association, the more likely it is that they form a word. The formula used for calculating the association score for adjacent two characters is:

$$A(xy) = MI(x, y) = log_2(\frac{\frac{freq(xy)}{N}}{\frac{freq(x)}{N}\frac{freq(y)}{N}}) \simeq log_2(\frac{p(xy)}{p(x)p(y)}) \quad (1)$$

Here, $A(xy)$ is the association score of bi-gram characters xy; $freq(x)$ is the frequency of character x occurring in the given corpus; $freq(xy)$ is the frequency of two characters sequence (x followed by y) occurring in the corpus; *N* is the size, in characters, of the given corpus; p(x) is an estimate of the probability of character x occurring in corpus, calculated as $freq(x)/N$.

Based on Sproat & Shih's work, Dai et al. [2] further developed an improved mutual information(IMI) formula to segment bi-gram words using regression analysis:

$$Improved\,MI(xy) = 0.39 * log_2(p(xy)) - 0.28 * log_2(p(x)) - 0.23 * log_2(p(y)) - 0.32 \quad (2)$$

Their experiment results indicate using this formula has similar precision with that of original mutual information formula. They also developed another formula called contextual information(CI) formula which considers the frequency of the character preceding and the character following the bi-gram as well. Given a character sequence - $vxyz$, the association strength of bi-gram xy is calculated from:

$$\begin{aligned} CI(xy) = {} & 0.35 * log_2(p(xy)) + 0.37 * log_2(p(v)) \\ & + 0.32 * log_2(p(z)) - 0.36 * log_2(p_{docwt}(vx)) \\ & - 0.29 * log_2(p_{docwt}(yz)) + 5.91 \end{aligned} \quad (3)$$

Where p_{docwt} is the weighted probability for given the character or bi-gram in corpus by considering frequency of document where that character or bi-gram appears. The contextual information formula has been proven better in term of precision. There is a 7% improvement in average comparing with IMI formula.

3 N-Gram Mutual Information

To overcome the limitations of the mutual information approaches including its extensions IMI and CI in recognising words with two characters only, we propose a new simple unsupervised method - n-gram mutual information (NGMI) to segment n-gram words. Phrase mutual information is developed based on mutual information of segments by expanding it with contextual information. The idea is to search words by looking for the word boundaries inside a given sentence by combining contextual information, rather than looking for words. This was tried by Sun et al. [9]before. The two adjacent characters are ``bounded" or ``separated" through a series of judgment rules based on values of mutual information and difference of t-score. But for NGMI there are no rules involved, and mutual information of **segments** not just adjacent characters is considered. The boundary | of a sub-string $(L|R)$, consisting of a left substring L, and a right substring R, is determined based on the boundary confidence(BC). BC measures the association level of the left and right substrings. The Boundary Confidence of any adjoining segments is defined as:

$$\begin{aligned} BC(L|R) &= MI(L, R) \\ &= sgn * (A(LR))^2 \end{aligned} \quad (4)$$

Where,

$$sgn = \begin{cases} -1, & if\ A(LR) < 0 \\ 1, & if\ A(LR) >= 0 \end{cases}$$

Here, A is the association score of segment S_i and segment S_{i+1}. The lower the mutual information score of *L* and *R*, the more confident we are about the boundary. Generally speaking, characters that occur together frequently have a high mutual information value, indicating a strong association between them; it is then unlikely that there will be a boundary between them. When the boundaries are determined, the characters between the boundaries are considered as candidate words.

For any input string, we have

$$s = c_1c_2c_3 \cdots c_ic_{i+1} \cdots c_n \quad (5)$$

Here, s is the a input string - containing n Chinese characters. There may be a boundary between any pair of adjoining characters c_ic_{i+1}. Given a sequence of n characters we can derive a complete list of all possible segmentations. So for each possible segmentation S, we have

$$\begin{aligned} S &= [c_1c_2 \cdots c_i] \,|\, [c_{i+1}c_{i+2} \cdots c_{i+k}] \,|\, \cdots \,|\, [c_{n-m}c_{n-m+1} \cdots c_n] \\ &= S_1S_2 \cdots S_x \end{aligned} \quad (6)$$

$[c_1c_2c_3 \cdots c_l]$, or S_i,is a single segment from the entire sequences, a candidate word. Whether a certain segmentation has the correct words selected needs to be decided by a model that can make the best choice based on the ranking scores for all possible segmentations. These scores are calculated by accumulating all boundary confidence values. The n-gram mutual information formula is then defined as:

$$\begin{aligned} NGMI(S) &= [BC(S_1|S_2), BC(S_2|S_3), \cdots, BC(S_{n-1}|S_n)] \\ &= \sum_{i=1}^{n} BC(S_i|S_{i+1}) \\ &= \sum_{i=1}^{n} MI(S_i, S_{i+1}) \end{aligned} \quad (7)$$

In previous work by Sproat & Shih [12] and Dai et al. [2], the mutual information was only used to deal with two characters at a time. The N-Gram Mutual Information overcomes this limitation; it is using the mutual information in a manner which is different on two important counts. Firstly, by detecting boundaries the length of the words between adjacent boundaries are of variable lengths, and secondly, by looking at the segmentation of multiple words at once rather than one word at a time. In this paper, boundary confidence is calculated in a few varieties:

MI_{pair}, MI_{sum}, MI_{min}, MI_{max} *and* MI_{mean}.

Providing that the length of sub-string S_i is *n*, and the length of sub-string S_{i+1} is *m*, we have,

$$\begin{aligned} MI_{pair}(S_i, S_{i+1}) &= MI(C_{rightmost}(S_i), C_{leftmost}(S_{i+1})) \\ &= MI(C_n(S_i), C_1(S_{i+1})) \end{aligned} \quad (8)$$

Here, $C_n(S_i)$ or $C_{rightmost}(S_i)$ is the right most character of S_i , $C_1(S_{i+1})$ or $C_{leftmost}(S_{i+1})$ is the left most character of S_{i+1}. If considering only at most two characters each side of the boundary, we have

$$\begin{aligned} MI_{sum}(S_i, S_{i+1}) &= MI(C_n(S_i), C_1(S_{i+1})C_2(S_{i+1})) \\ &+ MI(C_n(S_i), C_1(S_{i+1})) \\ &+ MI(C_{n-1}(S_i)C_n(S_i), C_1(S_{i+1})) \\ &+ MI(C_{n-1}(S_i)C_n(S_i), C_1(S_{i+1})C_2(S_{i+1})) \end{aligned} \quad (9)$$

$$\begin{aligned} MI_{min}(S_i, S_{i+1}) &= min(MI(C_n(S_i), C_1(S_{i+1})C_2(S_{i+1})), \\ &MI(C_n(S_i), C_1(S_{i+1})), \\ &MI(C_{n-1}(S_i)C_n(S_i), C_1(S_{i+1})), \\ &MI(C_{n-1}(S_i)C_n(S_i), C_1(S_{i+1})C_2(S_{i+1}))) \end{aligned} \quad (10)$$

$$\begin{aligned} MI_{max}(S_i, S_{i+1}) &= max(MI(C_n(S_i), C_1(S_{i+1})C_2(S_{i+1})), \\ &MI(C_n(S_i), C_1(S_{i+1})), \\ &MI(C_{n-1}(S_i)C_n(S_i), C_1(S_{i+1})), \\ &MI(C_{n-1}(S_i)C_n(S_i), C_1(S_{i+1})C_2(S_{i+1}))) \end{aligned} \quad (11)$$

$$\begin{aligned} MI_{mean}(S_i, S_{i+1}) &= (MI(C_n(S_i), C_1(S_{i+1})C_2(S_{i+1})) \\ &+ MI(C_n(S_i), C_1(S_{i+1})) \\ &+ MI(C_{n-1}(S_i)C_n(S_i), C_1(S_{i+1})) \\ &+ MI(C_{n-1}(S_i)C_n(S_i), C_1(S_{i+1})C_2(S_{i+1})))/k \end{aligned} \quad (12)$$

Here, $C_{n-1}(S_i)$ is the second character, if it exists, counting backward starting from boundary and the right most character of the left hand side sub-string S_i, $C_2(S_{i+1})$ is the second character, if it exists, counting forward starting from boundary and the left most character of the right hand side sub-string S_{i+1},

$$k = \begin{cases} 2, & length(S_i \, or \, S_{i+1}) <= 2 \, and \, at \, begining \, or \, end \, of \, S \\ 4, & length(S_i \, and \, S_{i+1}) > 2 \end{cases}$$

So

$$NGMI_{pair}(S) = \sum_{i=1}^{n-1} MI_{pair}(S_i, S_{i+1}) \quad (13)$$

$$NGMI_{sum}(S) = \sum_{i=1}^{n-1} MI_{sum}(S_i, S_{i+1}) \quad (14)$$

$$NGMI_{min}(S) = \sum_{i=1}^{n-1} MI_{min}(S_i, S_{i+1}) \quad (15)$$

$$NGMI_{max}(S) = \sum_{i=1}^{n-1} MI_{max}(S_i, S_{i+1}) \quad (16)$$

$$NGMI_{mean}(S) = \sum_{i=1}^{n-1} MI_{mean}(S_i, S_{i+1}) \quad (17)$$

Given overall scores of NGMI(*S*) for all possible segmentations, the lower the score of a segmentation, the more likely for it to have the right splits. For any particular split, if the boundary confidence MI values are negative, we are pretty confident that we are not splitting words in the middle. A detailed example may help explain this, given a short segmentation

$$S_1(ab|cdef) = [MI(ab, \, cdef)]$$

$$S_2(ab|c|def) = [MI(ab, \, c), \; MI(c, \, def)]$$

and calculate $NGMI_{min}(S_1)$ and $NGMI_{min}(S_2)$,

$$\begin{aligned} NGMI_{min}(S_1) = min(&MI(b, \, c), \; MI(ab, \, c), \\ &MI(b, \, cd), \; MI(ab, \, cd)) \end{aligned}$$

$$\begin{aligned} NGMI_{min}(S_2) = min(&MI(b, \, c), \; MI(ab, \, c), \\ &MI(b, \, cd), \; MI(ab, \, cd)) \\ &+ min(MI(c, \, d), \; MI(bc, \, d), \\ &MI(c, \, de), \; MI(bc, \, de)) \end{aligned}$$

4 Test Data

4.1 In-house Test Data

The following articles were chosen from the Chinese version of the Wikipedia: 本草纲目 (Bencao Gangmu), 马可・波罗 (Marco Polo), 張仲景 (Zhang Zhongjing),贫民百万富翁 (Slumdog Millionaire), 网络评论员 (50 Cent Party), and 风水 (Feng shui). All text from the above might be a mix of classical Chinese, simplified and traditional Chinese, and Chinese language variants. These pages were arbitrarily chosen simply as test pages.

4.2 Bake-off 2005 Test Data

In the Second International Chinese Word Segmentation Bakeoff test set, there are four groups of data (each having training, testing and gold-standard) provided by Academia Sinica, City University of Hong Kong, Peking University and Microsoft Research respectively [11]. The gold-standard data is segmented text following the word specifications defined by the each corpus creator. Each data set contains one form of Chinese writing either simplified or traditional.

	in-house	AS	CU	PKU	MSR
L.G.	81.27%	76.50%	79.58%	78.60%	73.14%
H.G.	18.73%	23.50%	20.42%	21.40%	26.86%

Table 1: Percentage of lower-gram and higher-gram words in test data

4.3 N-Gram Words Statistics For Test Data

Definitions:
Lower-gram words : 1-gram and 2-gram words
Higher-gram words : N-gram words, N > 3.

Table 1 shows around 20% of n-gram words in most test data sets are higher-gram words, and the rest of them (~80%) are lower-gram words, except that the standard gold data from Microsoft Research has lowest percentage of lower-gram words, only 73.14%. These statistical data indicates that simply searching for bi-gram words could not satisfy the need for n-gram word segmentation.

5 Experimental Design

5.1 String Pattern Frequency Table

The statistical information for the Chinese language is obtained through text mining of the Chinese Wikipedia XML corpus [3]. There are 56,662 documents, 27,360,399 Chinese characters, and 11,464 unique Chinese characters in total. For any character sequence with length less than 12, their corresponding frequencies are recorded in a string pattern frequency table. Since the size of such a complete table is very large, only those string patterns appearing in the corpus more than 216 times are kept. 216 is arbitrarily chosen to ensure that the frequency table can fit into program memory.

5.2 Stop Words

We recognise that an extra character like a preposition or a postposition cannot be separated from the actual word because of the strong statistical association. From the string pattern frequency table, the top 20 single-character words with the highest frequency (over 100,000 times) were selected as ``stop words".

5.3 Segmentation Runs

The Chinese segmentation experiments were performed using different segmentation methods and test data. Their run names and descriptions are listed in Table 2:

6 Segmentation Algorithms

6.1 MI Algorithm

The algorithm used in MI run to segment bi-gram words is that of Sproat and Shih [12]. As they did, the bi-gram frequency table keeps those words with a frequency greater than 4, and the threshold is set to 2.5. Given an input string of characters, the association strengths of each pair of adjoining characters are looked up. The pair with highest value is picked, then the second highest. If there are pairs with the same values, the right most pair is chosen. The bi-gram word with the highest value amongst the rest is repeatedly chosen until no pair's score is higher than the threshold. The remaining characters are then considered as one-character words.

6.2 IMI Algorithm

In Dai et al. IMI method, two sets of different algorithms, Comparative Forward Match(CFM) and Forward Match(FM) were implemented to perform the segmentation [2]. The segmentation process for both algorithms starts from the beginning of the sentence, and continues until the end. CFM is slightly more precise than FM in all their experiments. In our IMI run, we use the CFM algorithm to segment bi-gram words. The bi-gram frequency table only keeps words with frequency higher than 4, and the threshold is set to -2.5, which is the parameter used by Dai et al. having a segmentation result with the highest recall and lowest precision rate in all their IMI experiments. Given the sentence ABCDE, for example, the steps of segmenting it with CFM algorithm are:

First only the bi-gram AB is considered. If the association score of it is lower than the threshold, A is a single character word. Then BC is next to be considered. However, if the score of AB is higher than the threshold, then both bi-grams AB and BC are considered. If BC also has a score above the threshold but AB's is higher, AB is then chosen as a word. On other hand, if BC has the higher value, A is then marked as a 1-character word and CD also needs to be considered to decide whether BC is a bi-gram word. This process repeats until all words are segmented.

6.3 NGMI Algorithm

Given a Chinese character sequence: $s = c_1c_2c_3 \cdots c_ic_{i+1} \cdots c_n$, the word segmentation process using the NGMI has following steps:

1. The first *x* characters (in the experiments, *x* was set to 11) : $c_1c_2c_3 \cdots c_x$ are retrieved from the unsegmented text *s* for segmenting.
2. Build a list of all possible segmentations - S_{list} of the *x* characters. The upper bound of S_{list}equals $2^x - 1$. For 11 characters, there will be 1023 permutations, but segmentations will be removed from the list if they contain any

Run Name	Description
ICTCLAS	With ICTCLAS Chinese word segmentation system online demonstration version [6], from Chinese Academy of Sciences, using the in-house test data. It was developed based on multi-layer hidden Markov model [5]
MI	With original mutual information formula [12]using the in-house test data
IMI	With the improved mutual information formula proposed by Dai et al. [2]using the in-house test data
NGMI_PAIR	With $NGMI_{pair}$ formula using the in-house test data
NGMI_SUM	With $NGMI_{sum}$ formula using the in-house test data
NGMI_MIN	With $NGMI_{min}$ formula using the in-house test data
NGMI_MAX	With $NGMI_{max}$ formula using the in-house test data
NGMI_MEAN	With $NGMI_{mean}$ formula using the in-house test data
NGMI_MIN_SW	With $NGMI_{min}$ formula combining stop-words judgment using the in-house test data. The segmentation process repeats on already segmented words with length more than two characters, the words will be further split if the conditions as stipulated are met, (see ``step 6" in the segmentation algorithm)
NGMI_MIN_SW_AS	Same with NGMI_MIN_SW run but using the Academia Sinica test data
NGMI_MIN_SW_CU	Same with NGMI_MIN_SW run but using the City University of Hong Kong test data
NGMI_MIN_SW_PU	Same with NGMI_MIN_SW run but using the Peking University test data
NGMI_MIN_SW_MSR	Same with NGMI_MIN_SW run but using the Microsoft Research test data

Table 2: The list of all segmentation runs

substring that is not in frequency table.

$$S_{list} = \begin{cases} c_1|c_2c_3\cdots c_x \\ c_1|c_2|c_3\cdots c_x \\ c_1|c_2c_3|\cdots c_x \\ \cdots \end{cases}$$

3. For each boundary in the candidate segmentation, apply boundary confidence calculation, and sum the BC scores for each segmentation.

4. Sort S_{list} based on the segmentation scores in ascending order. It means that the lower score, the more likely it is to have the correct boundaries.

5. Choose the first segmentation (having highest rank) as the best segmentation:

$$S_{best} = c_1c_2c_3\cdots c_x = W_1W_2\cdots W_y$$

6. This step will only be executed if this is a stop words elimination run (NGMI_MIN_SW or NGMI_MIN_SW_AS, etc.). For any word-$W_i(c_1c_2\cdots c_k)$ in the best segmentation S_{best} with length more than two characters, it will be further broken down as in previous four segmentation steps(step 2 to step 5) into:

$$W_i = w_{i1}w_{i2}\cdots w_{iz}$$

This further segmentation will be accepted only if it meets the following conditions: both the first segment (w_{i1}) and the last segment (w_{iz}) of W_i are not one-character word; or if either w_{i1} or $w_{iz}$is a one-character word and it is in stop words list. For example, if W_2contains a stop word w_{21} at the beginning, then the best segmentation S_{best} from step 5 now become:

$$S_{best} = W_1[w_{22}w_{23}\ldots w_{2z}]\cdots W_y$$

7. Accept all the segments of S_{best} as words except for putting the last word W_y back into the unsegmented text. The last segmented word is returned to the unsegmented text since the split of these *x* characters was arbitrary and the best segmentation S_{best}may have split a long word in the middle.

8. Start the segmentation loop and repeat the segmentation process from step 1-7, until all the remaining characters are consumed.

The current version of NGMI algorithm isn't optimised yet. The segmentation performance is approximately 6000 words per second.

7 Evaluation and Analysis

In this section we compare the performance of the different segmentation runs on both the in-house test data and the bake-off test data. The comparison of the precisions in the in-house test data is used as major performance measurement for identifying the best NGMI variant. Also, the performance of segmentation runs on the all data is measured by overall recall rate.

7.1 Runs On the In-house Test Data

The precision values and their corresponding numbers of correctly identified words in each run against in-house test data are

given in Table 3. The recall figures of all runs on the in-house test data are given in Table 4.

Table 3 shows the mutual information runs are inherently limited by selecting only bi-gram words; and the NGMI runs are able to extract words with up to seven characters, even though the NGMI runs achieve only around 50% precision rate overall. The number of correctly identified words are similar for all runs on the in-house test data. The mutual information runs identify a high number of bi-gram words accurately, and the NGMI_MIN_SW run produces similar results but with more higher-gram words correctly identified. The results of the NGMI_MIN_SW run demonstrate an increase in the overall precision rate, reaching 62.64% from 53.52%, but the numbers of correctly identified higher-gram words drop. Some of the correct n-gram words are split and lost due to the further segmentation.

Despite the loss of correctly identified n-gram words, the NGMI_MIN_SW run still has the highest recall rate of all runs. The recall rates of mutual information runs, 69.17% and 68.61% respectively, come second and third. Other NGMI runs have slightly over 60% recall rate. The recall rate of ICTCLAS(56.58%) is low considering its relatively high precision. And considering the number of the single character words identified by the ICTCLAS run on the in-house test data is significantly higher than those in other runs but with a low precision, this suggests that the ICTCLAS online word segmentation system is accurate at recognising one form of written Chinese(either simplified or traditional), but it fails in the other. In mixed form documents the use of ICTCLAS could be problematic.

Overall, the supervised methods normally restrict themselves to choose words from the lexicon only, so their segmentation results have relatively a small number of found words. This explains why ICTCLAS has a high precision but a low recall. In contrast, as there isn't a finite correct words set for NGMI runs, the number identified words could be huge. And that leads to the decrement in the precision because of the larger denominator.

n-gram	ICTCLAS		MI		IMI		NGMI_PAIR		NGMI_SUM		NGMI_MIN		NGMI_MAX		NGMI_MEAN		NGMI_MIN_SW	
1	36.35%	434	42.05%	410	42.87%	406	49.39%	365	48.95%	350	48.79%	362	48.65%	343	50.00%	342	49.40%	368
2	87.20%	1042	79.02%	1318	71.06%	1429	71.64%	1329	63.35%	1279	69.74%	1307	61.82%	1279	64.39%	1318	72.76%	1520
3	83.62%	97		0		0	18.49%	113	18.65%	124	19.51%	104	18.50%	123	19.16%	123	29.91%	96
4	75.68%	28		0		0	16.29%	29	13.51%	30	14.21%	27	13.30%	29	13.36%	29	33.33%	8
5	100.00%	3		0		0	9.09%	2	6.25%	2	4.76%	1	6.06%	2	6.06%	2	0.00%	0
6		0		0		0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0		0
7		0		0		0	0.00%	0	0.00%	0	50.00%	1	0.00%	0	0.00%	0		0
Overall	63.03%	1604	65.38%	1728	62.04%	1835	53.88%	1838	48.72%	1785	53.52%	1802	48.00%	1776	49.93%	1814	62.64%	1992

Table 3: The precision and corresponding number of correctly identified words in the in-house test data. Overall precision = # of correctly identified words / # of all detected words.

7.2 Runs On the Bake off Test Data

It has to be noted that all the segmentation runs on the bake-off test data are created directly using the string frequency table obtained from the Chinese Wikipedia corpus without any knowledge of the bake-off training data. The training text is in fact completely independent of the test corpus. The recall figures for all bake-off runs are given in Table 5.

Table 5 shows that the recall rates of all bake-off run are around 70%, which indicates the corpus independent ability of NGMI in segmenting n-gram words. Of course this can be attributed to the fact that the Chinese Wikipedia corpus is a mixed language corpus and hence it covers the language of the bake-off text. Table 4 and table 5 also show that the recall rate of NGMI method using $NGMI_{min}$ formula with stop words elimination is the highest, and consistent (all around

ICTCLAS	MI	IMI	NGMI_PAIR	NGMI_SUM	NGMI_MIN	NGMI_MAX	NGMI_MEAN	NGMI_MIN_SW
56.58%	60.95%	64.73%	64.83%	62.96%	63.56%	62.65%	63.99%	70.26%

Table 4: Recall rate of segmentation runs using in-house test data

NGMI_MIN_SW_AS	NGMI_MIN_SW_CU	NGMI_MIN_SW_PU	NGMI_MIN_SW_MSR
68.96%	72.85%	72.26%	69.86%

Table 5: Recall of segmentation runs on the bake-off test data

70%) through all the runs.

8 Conclusions

In this paper, we have presented a simple unsupervised method NGMI using purely the Chinese text statistics drawn from the Wikipedia corpus to segment n-gram words. It is based on mutual information theory, but overcomes the limitation of the original mutual information based methods in recognising only bi-gram words by introducing the judgment of boundary-confidence of the adjacent segments.

To examine the feasibility of segmentation with n-gram mutual information and to find the best n-gram mutual information formula, a set of segmentation runs including a run using a state-of-the-art word segmentation system (ICTCLAS), two runs using different mutual information formulas (MI and IMI) and five runs using different n-gram mutual information variants ($NGMI_{pair}$, $NGMI_{sum}$, $NGMI_{min}$, $NGMI_{max}$, and $NGMI_{mean}$) were produced for performance comparison. Our experiments show $NGMI_{min}$ method performed best among all variants. The precision, number of correctly identified words, and overall recall rate of NGMI segmentation runs show encouraging results in segmenting n-gram words for Chinese Wikipedia articles.

As NGMI is a simple unsupervised method without needing much knowledge of the language, it will certainly benefit the text processing, when segmentation is required and situations are new to the-state-of-the-art systems, by providing the baseline n-gram word segmentation.

References

[1] Second international chinese word segmentation bakeoff - result summary. `http://www.sighan.org/bakeoff2005/data/results.php.htm`.

[2] Dai, Y., Loh, T. E., and Khoo, C. S. G. A new statistical formula for Chinese text segmentation incorporating contextual information. 82--89.

[3] Denoyer, L., and Gallinari, P. The Wikipedia XML Corpus. Tech. rep.

[4] Haodi Feng, Kang Chen, C. K. X. D. *Unsupervised Segmentation of Chinese Corpus Using Accessor*. Springer Berlin / Heidelberg, 2005, pp. 694--703.

[5] Institute of Computing Technology, Chinese Academy of Sciences. Chinese lexical analysis system ICTCLAS. `http://www.ict.ac.cn/jszy/jsxk_zlxk/mfxk/200706/t20070628_2121143.html`.

[6] Institute of Computing Technology, Chinese Academy of Sciences. ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System). `http://ictclas.org`.

[7] Lua, K. From character to word - An application of information theory. *Computer Processing of Chinese & Oriental Languages 4*, 4 (1990), 304--312.

[8] Lua, K., and Gan, G. An application of information theory in Chinese word segmentation. *Computer Processing of Chinese & Oriental Languages 8*, 1 (1994), 115--124.

[9] Maosong, S., Dayang, S., and Tsou, B. K. Chinese word segmentation without using lexicon and hand-crafted training data. In *Proceedings of the 17th international conference on Computational linguistics* (Morristown, NJ, USA, 1998), Association for Computational Linguistics, pp. 1265--1271.

[10] Peng, F., Feng, F., and McCallum, A. Chinese segmentation and new word detection using conditional random fields. 562.

[11] SIGHAN. Second International Chinese Word Segmentation Bakeoff Data. `http://www.sighan.org/bakeoff2005/`.

[12] Sproat, R., and Shih, C. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese & Oriental Languages 4*, 4 (1990), 336--351.

[13] Teahan, W. J., McNab, R., Wen, Y., and Witten, I. H. A compression-based algorithm for Chinese word segmentation. *Comput. Linguist. 26*, 3 (2000), 375--393.

[14] Tung, C.-H., and Lee, H.-J. Identification of unknown words from a corpus. *Computer Processing of Chinese & Oriental Languages 8*, 1 (1994), 115--124.