A Lexicon-Corpus-based Unsupervised Chinese Word Segmentation Approach

Lu Pengyu, Pu Jingchuan, Du Mingming, Lou Xiaojuan and Jin Lijun
School of Management, Harbin Institute of Technology, Harbin, China
Email: lupengyu@hit.edu.cn

Submitted:  Accepted:  Published:

Abstract- This paper presents a Lexicon-Corpus-based Unsupervised (LCU) Chinese word segmentation approach to improve the Chinese word segmentation result. Specifically, it combines advantages of lexicon-based approach and Corpus-based approach to identify out-of-vocabulary (OOV) words and guarantee segmentation consistency of the actual words in texts as well. In addition, a Forward Maximum Fixed-count Segmentation (FMFS) algorithm is developed to identify phrases in texts at first. Detailed rules and experiment results of LCU are presented, too. Compared with lexicon-based approach or corpus-based approach, LCU approach makes a great improvement in Chinese word segmentation, especially for identifying n-char words. And also, two evaluation indexes are proposed to describe the effectiveness in extracting phrases, one is segmentation rate (S), and the other is segmentation consistency degree (D).

Index terms: Chinese word segmentation, lexicon-based, Corpus-based, word frequency, natural language processing.
I. INTRODUCTION

With the dramatic increase of User Generated Content (UGC), the theory and methodology of unstructured text analysis and mining play a more important role in areas of knowledge management, business intelligence, management decision-making, social public opinion analysis, etc. At the same time, to deal with information intelligently is highly desired for automation industry [1]. However, the current generation of algorithms for feature extraction and recognition seems to have somewhat reached their limits [2]. Fortunately, as an essential step in Chinese text analysis and mining, Chinese word segmentation can also significantly impact the results of text data mining [3; 4; 5]. Therefore, how to segment words effectively is a hot issue in recent years.

To improve the performance, scholars have been devoting much energy and improving various methods to tackle two main causes of segmentation errors, namely, segmentation ambiguities and unknown (or out-of-vocabulary, OOV) words [6]. The basic technique for identifying distinct words from Chinese texts is the lexicon-based identification approach which works by comparing strings with the words in lexicon, and it has higher technical maturity than other approaches based on machine learning or statistics [7], but it cannot perform Chinese word segmentation process very well, because many OOV words cannot be identified by a constant lexicon [8; 9], such as persons’ name, places’ name and highly changeable texts with time, etc. So the extraction of OOV words becomes a key technology for the lexicon-based Chinese word segmentation system [10; 11; 12; 13; 14; 15]. Moreover, poor word segmentation results usually occur when a general lexicon is used in a lexicon-based Chinese word segmentation system to segment texts in specific domains, and compared to segmentation ambiguities, OOV words account for more errors [16]. The best solution is to detect new words from the corpora of specific domain [17; 18; 19; 20]. It is reported that segmentation accuracy is significantly higher when the lexicon is constructed by the same type of corpus, which is used for test [21]. Corpus-based approach is effective in identifying OOV words, but it has the disadvantages of inevitably corrupting many known words. Up to now, a mature segmentation system generally combines different kinds of
algorithms, in order to improve the precision and accelerate the speed of segmenting word at the same time. Therefore, this paper proposes a Lexicon-Corpus-based Unsupervised (LCU) Chinese word segmentation approach, which combines the lexicon-based approach with the corpus-based approach in order to take advantages of each approach. Conventionally, before matching strings with words in the corpus or lexicon we should segment texts by using algorithm, such as Forward Maximum Matching (FMM) and Reverse Maximum Matching (RMM), but they all have inevitable shortcoming that they cannot guarantee all phrases extracted because they extract words from beginning or end of text rigorously, so we propose an algorithm which extracts phrases from texts in priority called Forward Maximum Fixed-count Segmentation (FMFS). Experiments demonstrate that this approach is more effective than the lexicon-based approach or the corpus-based one on Chinese word segmentation, especially for OOV words. By using FMFS, the performance of extracting phrases is highly improved.

This paper is organized as follows. After a general introduction of word segmentation approaches, an overview of LCU approach is proposed in Section II. A method for texts processing is described in section III. FMFS algorithm is illustrated in Section IV. Word segmentation rules, which are the key part of the approach, are presented in detail in Section V. Experiment results and evaluations are discussed in Section VI. Conclusions, including a summary of our major contributions and an outlook for future work, are stated in Section VII.

II. OVERVIEW OF THE APPROACH

A Chinese word can be considered as a stable string of characters. In this paper, n-char stands for a string of $n$ consecutive Chinese characters. If $n = 1$, that means there is only one Chinese character in the word, we call it uni-char word. Similarly, $n = 2, 3, 4$, we call it bi-char word, tri-char word, and quad-char word respectively. Moreover, if $n > 4$, we call it n-char word directly.

The aim of this paper is to provide an approach, which can extract words from these strings effectively, especially for n-char words. So n-char words will be extracted in priority by this
approach. In order to achieve the aim, we develop a Forward Maximum Fixed-count Segmentation (FMFS) algorithm to get strings with the same maximum characters count \( n \), which we predefined at the same time. Then, scanning the lexicon firstly will identify words. If there are some strings, which can’t be identified by lexicon, they will be continuously identified by scanning corpus. Afterwards, if there are still some strings remained, let \( n = n - 1 \), then repeat the procedures mentioned above. Algorithm will end when \( n < 2 \) or all the strings have been segmented into words. Figure 1 illustrates the overview of the approach.

**Figure 1** Overview of LCU Chinese Word Segmentation Approach

### III. TEXT PREPROCESSING

The objective of text preprocessing is to get rid of punctuation characters and stop words in the original text. First, put the original Chinese text into set \( S \) as (1).

\[
S = \{c_1c_2 \cdots c_w, c_{w+1} \cdots c_z\}
\]  

(1)

Second, get the first Chinese character string from the text beginning to the first non-Chinese character (punctuation character, spaces, etc.) and put the string into another set \( C \) to be segmented.

\[
C = \{c_1c_2 \cdots c_w\}
\]  

(2)
Where \( w \) is the count of Chinese characters in pre-segmented word set \( C \). After that we remove the string from \( S \), which means \( S = S - C = \{ c_{w+1} \cdots c_z \} \). Third, extract the first non-Chinese character of set \( S \), put it into word segmentation result set \( C' \), and remove it from \( S \). Last, extract stop words from \( C \) and put them into \( C' \), the stop words in \( C \) are replaced by “//”. Supposing \( c_{p+1} \) is a stop word, then

\[
S = \{ c_{w+1} \cdots c_z \} \\
C' = \{ c_{p+1} \} \\
C = \{ c_1 c_2 \cdots c_p // c_{p+2} c_{p+3} \cdots c_w \}
\]

IV. FORWARD FIXED-COUNT SEGMENTATION ALGORITHM

Forward Maximum Fixed-count Segmentation (FMFS) algorithm is developed to get all n-char strings from a character string in set \( C \). First, Extract the character string before the first symbol “//” in set \( C \) as (5) (if there is no “//” in set \( C \), then extract the whole string in set \( C \).) and put it into set \( W \). Then

\[
W = \{ c_1 c_2 \cdots c_p \}
\]

Let \( n \) be the maximum count of extracted words’ character, and \( 2 \leq n \leq p \). Then the string in \( W \) will be segmented into several n-char strings from left to right in turns, just like

\[
w_{n_1} = c_1 c_2 c_3 \cdots c_n \quad w_{n_2} = c_2 c_3 c_4 \cdots c_{n+1} \quad \cdots \quad w_{n_{p-n+1}} = c_{p-n+1} c_{p-n+2} \cdots c_p
\]

Put these segmented strings into candidate words set \( W' \). Generally, a Chinese word consists of two or more characters, and most words consist of less than eight characters. So \( n \) ranges from 2 to 8 in this paper. Then

\[
W' = \{ w_{n_k} \mid w_{n_k} = c_k c_{k+1} \cdots c_{k+n-1}, \quad n \in [2,8], \quad k \in [1, p-n+1] \}
\]

Where \( k \) is the sequence number of characters in set \( W \), \( w_{n_k} \) is the \( k^{th} \) segmented string in \( W' \) which contains \( n \) continuous characters. The amount of segmented strings is \( p - n + 1 \). For instance, supposing that \( p = 10 \), then \( W = \{ c_1 c_2 \cdots c_{10} \} \). Let maximum character count
n = 6, then the amount of segmented strings is \( p - n + 1 = 10 - 6 + 1 = 5 \). These strings are

\[ w_6 = \{c_1, c_2, \ldots, c_6\}, \quad w_5 = \{c_2, c_3, \ldots, c_7\}, \quad w_4 = \{c_3, c_4, \ldots, c_8\}, \quad w_3 = \{c_4, c_5, \ldots, c_9\}, \quad w_2 = \{c_5, c_6, \ldots, c_{10}\} \].

In the same way, if \( n = 5 \), there are 6 segmented strings, more specifically,

\[ w_5 = \{c_1, c_2, \ldots, c_5\}, \quad w_4 = \{c_2, c_3, \ldots, c_6\}, \quad w_3 = \{c_3, c_4, \ldots, c_7\}, \quad w_2 = \{c_4, c_5, \ldots, c_8\}, \quad w_1 = \{c_5, c_6, \ldots, c_9\}, \quad w_0 = \{c_6, c_7, \ldots, c_{10}\}. \]

V. WORD SEGMENTATION RULES

**Rule 1. Lexicon-based approach has priority to identify words over corpus-based approach.**

After getting candidate words set \( W' \) like \( (7) \), \( w_n \) in \( W' \) will be identified as segmented words by scanning the lexicon at first. If there is a word in lexicon, which can match with \( w_n \) perfectly, put the continuous characters \( c_k c_{k+1} \cdots c_{k+n-1} \) in \( w_n \) into word segmentation result set \( C' \), and remove the character series \( c_k c_{k+1} \cdots c_{k+n-1} \) in \( w_n \) from \( W \). Then

\[
C' = \{c_k c_{k+1} \cdots c_{k+n-1} / c_{p+1} // / \} \tag{8}
\]

\[
W = \{c_1 \cdots c_{k-1} / c_{k+1} \cdots c_p \} \tag{9}
\]

If there is no word in lexicon which can match with \( w_n \) perfectly, then the corpus-based approach will be used to identify whether \( w_n \) is a word or not.

When corpus is large enough, words with some meanings must occur repeatedly in it. So a word could be identified by its occurrence frequency in corpus. The occurrence frequency in corpus is called corpus type frequency in this paper. The minimum corpus type frequency that a word could be identified is called threshold. Let \( t_n \) represent the threshold of n-char words, \( f_n \) be the corpus type frequency of an n-char string. If \( f_n \geq t_n \), it means that the n-char string could be identified as a n-char Chinese word. A data set of threshold values \( (T) \) is formed as \( (10) \).
Let \( f_{w_k} \) be the corpus type frequency of the \( k^{th} \) n-char words in \( W' \), the data set \( (f_{w_k}) \) of corpus type frequency for n-char words in \( W' \) can be described as (11).

\[
f_{w_k} = \{ f_{w_{k_1}}, f_{w_{k_2}}, \cdots, f_{w_{k_n}} \}
\]

Where \( k \in [1, p-n+1] \), \( n \in [2, 8] \). If \( f_{w_{k_i}} \geq t_n \), then put the character string \( c_k c_{k+1} \cdots c_{k+n-1} \) of \( w_{k_i} \) into \( C' \), remove the character string \( c_k c_{k+1} \cdots c_{k+n-1} \) of \( w_{k_i} \) from \( W \). Result is same as (8)(9).

**Rule 2.** The word with maximum \( n \) has priority to be identified.

Generally, phrase is better than word to express textual meaning. That is to say, we should try our best to identify phrase as a basic unit to express the textual meaning perfectly. There are some statistic approaches developed to solve this problem, such as mutual information, n-gram statistic model, \( t \)-testing principle, etc. Almost all the statistic approaches are based on the concurrence frequency of the adjacent words. In fact, if there is a high concurrence frequency of the adjacent words, as a continuous character string, these adjacent words must have a high corpus type frequency. Thus, we can identify a phrase as an n-char word by corpus type frequency directly.

In common conditions, there are more characters in a phrase than in a word. So if the words with maximum \( n \) are identified at first, the probability of identifying phrases will be enhanced. So the LCU word segmentation approach specifies that the words with maximum \( n \) have priority to be identified. Therefore, FMFS is used to segment strings, and ensure the word with maximum \( n \) to be identified in priority. By this rule, even though \( c_i c_{i+1} \) and \( c_{i+2} c_{i+3} \) could be identified as words separately based on type frequency in lexicon or corpus, they will not be put into the result set \( C' \) because \( c_i c_{i+1} c_{i+2} c_{i+3} \) will be identified as a word firstly. This rule is able to keep the words’ semantic integrality.
Rule 3. The word with higher frequency in lexicon or corpus takes priority to be identified.

If there is more than one word with the same part of a character string meeting the rule at the same time during the word segmentation by Rule 1, which word could be put into the result set $C'$ is decided by the maximum word frequency in lexicon or corpus. For example, there are two n-char strings, $w_{n_i} = c_{i1}c_{i2}\cdots c_{in}$, and $w_{n_{i+1}} = c_{i1}c_{i2}\cdots c_{in}n$, and $f_{w_{n_i}} > f_{w_{n_{i+1}}}$, $f_{w_{n_{i+1}}} \geq f_{w_{n_i}}$.

By Rule 1, $w_{n_i}$ and $w_{n_{i+1}}$ should be identified as words together, but there is a shared part $c_{i+1}\cdots c_{in}$, between these two strings. This part shouldn’t be identified as a word more than once. So which string will be identified as a word is confirmed by $\max\{f_{w_{n_i}}, f_{w_{n_{i+1}}}\}$.

Supposing that $\max\{f_{w_{n_i}}, f_{w_{n_{i+1}}}\} = f_{w_{n_i}}$, the n-char string $c_{i1}c_{i2}\cdots c_{in}$ in $w_{n_i}$ will be put into $C'$ as an identified word.

Rule 4. If there are one more way to segment words with the same $n$ from a character string, the way with maximum word count will be taken.

Sometimes, there are several ways to segment words from a character string. For example, bi-char words are to be extracted from a character string $c_1c_2c_3c_4$. There are two schemes, one makes the string be segmented to $c_1c_2//c_3c_4$, and the other makes the string be segmented to $c_1//c_2c_3//c_4$. Obviously there are two bi-char words in the first one and only one bi-char word in the second one. In this situation, the first scheme will be adopted.

Rule 5. If there are still some residue characters after bi-char word segmentation, put the residues into final result set directly.

After extracting bi-char words from a character string, maybe there are still some residue characters. The residue could contain one character, two characters, three characters, or even more. If there is only one character, it is usually a uni-char word. If there are more than one characters remained, they can probably constitute an unknown word, such as a person name, place name, or one terminology in a special domain. These words could not be identified
directly because of the imperfections of lexicon or corpus. So put them into the final result set directly. For instance, \( C' \) in (8) is the initial word segmentation result, and the strings in (9) cannot be extracted out any other words by lexicon-based or corpus-based approach. So \( c_1 \cdots c_{k-1} \) and \( c_{k+1} \cdots c_p \) in \( W \) are residues. At this time, put the two residue strings into \( C' \) as a result. So the final word segmentation result is \( C' = \{ c_1 \cdots c_{k-1} \} \cup \{ c_k c_{k+1} \cdots c_{k+n-1} \} \cup \{ c_{k+1} \cdots c_p \} \cup \{ c_{p+1} \} \}. \) This rule is beneficial for figuring out unknown words.

VI. EVALUATION OF EXPERIMENT RESULTS

a. Experiments Design
In order to measure the effectiveness of LCU word segmentation approach from different points of view, we carried out experiments with different approaches respectively. We designed three experiments to test the text segmentation performance of lexicon-based approach, corpus-based approach that is described in reference [11] and the LCU approach presented in this paper separately. 1.56M texts from Chinese corpus developed by Fudan University are selected as the test corpus. These texts involve 10 fields, including literature, history, education, electronics, sports, transportation, economy, law, medicine, and politics, and 10 texts are chosen from each field. Meanwhile, an online corpus developed by Chinese Language Research Center of Peking University (CCL corpus, http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp?dir=xiandai) is selected to get the corpus type frequency of character strings which are to be identified as words. The scale of CCL corpus in modern Chinese is 695MB. The lexicon used in experiments is approximately 837KB, 7891 words in all. Among the words in the lexicon, there are 1240 uni-char words, 5293 bi-char words, 915 tri-char words, 373 quad-char words, 37 5-char words, 29 6-char words, and 4 7-char words. Maximum \( n \) is 8 in experiments, and the set of threshold values is as follows.

\[
T = \{ t_2, t_3, \cdots, t_7, t_8 \} = \{ 2000, 1000, 500, 100, 20, 10, 8 \}
\]  

(12)

Among all tests conducted below, Precision rate \((P)\), Recall rate \((R)\), and the balanced F-score \((F)\) are used to evaluate the segmentation results. They are calculated as below.
Lu Pengyu, Pu Jingchuan, Du Mingming, Lou Xiaojuan and Jin Lijun, A LEXICON-CORPUS-BASED UNSUPERVISED CHINESE WORD SEGMENTATION APPROACH

\[ P = \frac{a}{a+b} \times 100\% \]  
\[ R = \frac{a}{c} \times 100\% \]  
\[ F = \frac{2RP}{P+R} \]

Where, \( a \) represents the count of accuracy words identified by machine, \( b \) represents the count of the inaccuracy words identified by machine, and \( c \) represents the actual words count in text.

b. Discussions of Experiment Results

b.1 Overall Word Segmentation Results by Three Approaches

The experiment results with three approaches are shown in Figure 2. It is observed that the \( R \) of lexicon-based approach is higher than that of corpus-based approach, and its \( P \) is lower than corpus-based approach’s. LCU approach developed in this paper which integrates the advantages of the lexicon-based one and corpus-based one gets the best results among three approaches, its \( R \), \( P \) and \( F \) are all higher than the other two approaches’. It indicates that the improvement in LCU is successful.
b.2 Word Segmentation Results with Different $n$ Values by Three Approaches

It is a difficult problem to be solved in identifying n-char words. So the effectiveness of three approaches in identifying with n-char words is compared as Figure 3, and their $R$, $P$ and $F$ are shown in Table 1.

<table>
<thead>
<tr>
<th>approaches</th>
<th>bi-char word</th>
<th>tri-char word</th>
<th>quad-char word</th>
<th>n-char word</th>
</tr>
</thead>
</table>
| $R$  
lexicon-based | 88.55% | 78.87% | 56.38% | 11.36% |
corpus-based | 82.88% | 56.34% | 71.81% | 61.36% |
LCU | 85.32% | 64.79% | 73.15% | 65.91% |
| $P$  
lexicon-based | 87.57% | 53.85% | 92.31% | 100.00% |
corpus-based | 92.17% | 50.63% | 72.79% | 45.76% |
LCU | 93.86% | 56.10% | 74.15% | 48.33% |
| $F$  
lexicon-based | 0.8806 | 0.64 | 0.7 | 0.2041 |
corpus-based | 0.8728 | 0.5338 | 0.7230 | 0.5243 |
LCU | 0.8940 | 0.6013 | 0.7365 | 0.5577 |

Figure 3 shows that $R$ decreases with the increase of $n$ generally. $R$ of bi-char words and tri-char words by lexicon-based approach are higher than the other two approaches, but its $R$ of quad-char words and n-char words are lower, especially for n-char words, only achieving 11.36%. $R$ of LCU approach is better than corpus-based approach in all. $R$ of tri-char words gets the lowest value by either corpus-based approach or LCU approach.
Lu Pengyu, Pu Jingchuan, Du Mingming, Lou Xiaojuan and Jin Lijun, A LEXICON-CORPUS-BASED UNSUPERVISED CHINESE WORD SEGMENTATION APPROACH

Figure 3 Comparisons of $R$ with Different $n$ Values

Figure 4 shows that $P$ of different $n$ values by LCU approach are all better than corpus-based approach. $P$ of bi-char words and tri-char words by LCU approach are the highest among three approaches, but $P$ of quad-char words and n-char words are lower than lexicon-based approach. For lexicon approach, $P$ tends to rise with the increase of $n$ except for bi-char words. Especially for n-char words, its $P$ achieves 100% in the experiment. In common, $P$ of tri-char words is the lowest one for each of approaches.

![Figure 4 Comparisons of $P$ with Different $n$ Values](image)

Figure 4 Comparisons of $P$ with Different $n$ Values

$F$ value is used to evaluate approaches comprehensively. Figure 5 shows that LCU approach is better than the other two. The bigger $n$ is, $F$ tends to be lower for all three approaches, except for $F$ of tri-char words.

To sum up, LCU approach is more effective than the other two, especially for n-char words, it is far more perfect than lexicon-based approach. What is worth to be noticed is that the segmentation result for tri-char words is not optimistic for every approach. After further analyzing, we find the reason lies in the fact that in lexicon-based approach, many quad-char words and n-char words are segmented into bi-char words or tri-char word, so the count of accuracy and inaccuracy words identified by machine is all higher, it gets a higher $R$ and a
lower $P$. In corpus-based approach and LCU approach, the improper threshold value results in more inaccuracy n-char words being identified, which are consisted of bi-char words and tri-char words actually. So the count of bi-char words and tri-char words being identified accurately is lower than reality. And also, some tri-char words are segmented into uni-char words and bi-char words because of the improper threshold value, the count of tri-char words being identified accurately is further decreased. However, this problem in LCU approach is improved better than in corpus-based approach because of its integration of lexicon-based approach.

b.3 Comparison of Word Segmentation Integrity of Three Approaches
Since phrase as an n-char word plays an important role in semantic analysis, it is significant to measure the word segmentation integrity for phrase. In this paper, we use n-char word segmentation rate ($\sigma$) shown as (16) to measure word segmentation integrity.

$$\sigma = \frac{1}{m} \quad (16)$$

Where $m$ is the words count, which an n-char word is segmented into. Supposing $c_1c_2c_3c_4c_5c_6$ is a 6-char word, if it is segmented into $c_1///c_2c_4c_5///c_6$, then $m = 3$, and
If it is segmented into $c_1c_2 \text{ }//\text{ }c_3c_4c_5c_6$, then $m = 2$, and $\sigma = \frac{1}{m} = \frac{1}{2}$. If it is identified correctly, then $m = 1$, and $\sigma = \frac{1}{m} = 1$. Furthermore, the word segmentation integrity of the whole text could be measured by the mean value of the segmentation rate ($S$) as below.

$$S = \frac{\sum_{i=1}^{N} \sigma_i}{q} \times 100\%$$

Where $\sigma_i$ represents the word segmentation rate of $i^{th}$ n-char in the whole text, and $q$ is the actual count of n-char words in the whole text. Obviously, the larger the word segmentation rate is, the better the word segmentation integrity becomes. When all n-char words can be perfectly identified, $S = 1$. The word segmentation rates $S$ by three approaches are shown in Figure 6. It could be concluded that in keeping word segmentation integrity, LCU approach is the best and lexicon-based approach is not satisfied.

In addition, according to the definition of $R$ and $P$, $R$ describes the ratio of accuracy words count identified by machine to the actual words count in text, and $P$ describes the ratio of accuracy words count identified by machine to the total words count identified by
machine. In the experiments, $R$ of n-char words by LCU approach is up to 65.91%, while its $P$ is only 48.33%. It means that more accuracy words are identified, and there are also more inaccuracy words among the segmented words at the time. That is to say, the words count identified by machine is more than the actual one in text. On the contrary, $R$ of n-char words by lexicon-based approach is only 11.36%, and its $P$ achieves 100%. It infers that only a few accuracy words are identified, but the identified ones are all accuracy words, and the words count identified by machine is less than the actual one in text. Therefore, when measuring the effectiveness of an approach, the consistency between words count identified by machine and the actual one in text should be considered, too. Segmentation consistency degree ($D$) shown as (10) is proposed in this paper to describe the consistency.

$$D = \frac{a + b}{c} - 1$$ \hspace{1cm} (18)

Where $D > 0$ means words count identified by machine is more than the actual one in text, $D < 0$ indicates words count identified by machine is less than the actual one in text, $D = 0$ demonstrates words count identified by machine is as well as the actual one in text. The greater $|D|$ is, the greater difference is between words count identified by machine and the actual one in text. Segmentation consistency degrees with different $n$ from the experiment of LCU approach are shown in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>bi-char word</th>
<th>tri-char word</th>
<th>quad-char word</th>
<th>n-char word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexicon-based</td>
<td>0.0113</td>
<td>0.4648</td>
<td>-0.3893</td>
<td>-0.8864</td>
</tr>
<tr>
<td>Corpus-based</td>
<td>-0.1008</td>
<td>0.1127</td>
<td>-0.0134</td>
<td>0.3409</td>
</tr>
<tr>
<td>LCU</td>
<td>-0.0910</td>
<td>0.1127</td>
<td>-0.0134</td>
<td>0.3636</td>
</tr>
</tbody>
</table>

Figure 7 shows that the segmentation consistency degrees of tri-char words, quad-char words and n-char words are not satisfied by lexicon-based approach. Being similar with the results of corpus-based approach, the results of LCU are better than the results of lexicon-based approach, especially for n-char words.
Influence Factors for Word Segmentation Results in LCU Approach

Though there is distinct improvement in LCU approach, some parts of requirements are not always satisfied. By carrying out an in-depth analysis of the results, we figure out that domain has a considerable influence on the word segmentation results. As is shown in Figure 8, it is clear that the result in art is obviously better than the result in law. After further observing CCL corpus we used in experiments, we found that 18.54% of texts CCL corpus are in art, but only 0.25% of texts are in law. And also, there are a lot of terminologies in law, transport, sports and other specific domains. It is difficult to identify the terminologies if there are less corpus texts in these domains. As a result, it is necessary to keep enough corpus texts in each domain to improve the word segmentation result.

Furthermore, experimental results reveal that stop words lists, spelling errors, special terms (such as person name, place name, etc.) and so on, also have great influence on word segmentation results. Moreover, threshold used to identify words by corpus type frequency is a key factor to affect the word segmentation results. So how to deal with stop words lists, spelling errors, special terms, and how to set up thresholds are the problems that need to be solved in the future.
VII. CONCLUSIONS

A lexicon-corpus-based unsupervised Chinese word segmentation approach is proposed in this paper. During word segmentation by LCU approach, lexicon-based approach is used to identify words at first, and then corpus-based approach is adopted if character strings could not be identified by lexicon-based approach. In order to keep the semantic integrity, the
Lu Pengyu, Pu Jingchuan, Du Mingming, Lou Xiaojuan and Jin Lijun, A LEXICON-CORPUS-BASED UNSUPERVISED CHINESE WORD SEGMENTATION APPROACH

n-char words with maximum \( n \) have priority to be identified. When there are several n-char strings satisfying the subjects to be identified as words, the one with the highest frequency in lexicon or corpus will be extracted. If there is one more way to segment a string into words, the way is selected by the maximum count of identified words. And the characters remained after identifying bi-char words will be put into result set directly. Experiment results indicate that comparing with lexicon-based approach and corpus-based approach, LCU approach makes great improvement in word segmentation, especially for identifying n-char words. However, there are still some works to do in the future, such as making the corpus richer in each domain, figuring out new approaches to enhance the effectiveness by dealing with stop words lists, spelling errors and special terms, or setting up proper thresholds.

ACKNOWLEDGEMENTS

The National Science Foundation in China (71071041), Nature Science Foundation in Hei Long Jiang Province of China (G201306), Natural Scientific Research Innovation Foundation in Harbin Institute of Technology (HIT.NSRIF.2009109), supported this research.

REFERENCES


