Unsupervised Learning and Linguistic Rule Based Algorithm for Uyghur Word Segmentation

Turdi Tohti and Winira Musajan
School of Information Science and Engineering, Xinjiang University, Urumqi, China
Email: turdy@xju.edu.cn, winira@xju.edu.cn

*Askar Hamdulla
School of Software Engineering, Xinjiang University, Urumqi, China
*Corresponding author, Email: askar@xju.edu.cn

Abstract—Inter-word spaces based traditional word segmentation method not very appropriate for multi-word structured semantic words due to the fact that it will split the semantic words into several fragments that inconsistent with its original meaning. So, this will be a bottleneck problem in Uyghur text analysis and text understanding applications. This paper puts forward a new idea and related algorithms for segmentation of Uyghur multiword structured semantic words. In this algorithm, the word based Bi-gram and contextual information are derived from large scale raw text corpus automatically, and according to the association rules between Uyghur words, the liner combinations of mutual information, difference of t-test and dual adjacent entropy are taken as a new measurement to estimate the agglutinative strength between two adjacent Uyghur words. The experimental result on large-scale open text shows that the proposed algorithm achieves 88.21% segmentation accuracy.

Index Terms—Uyghur Language; Semantic Word; Combined Statistics; Word Association Rule; Semantic Word Segmentation

I. INTRODUCTION

Word segmentation is the first step and also is a key step in natural language processing (NLP), what method to use and its difficulty is different in different language environment. But the ultimate goal is to get the minimum use of linguistic units that express specific independent semantics.

Uyghur language belongs to Turkish language group of Altaic language family, and also belongs to the agglutinating language on structure grammar, is a kind of alphabetic writing. Looking on the surface, Uyghur text is a word sequences that separated by inter-word spaces and on this feature is similar to English. For this reason, word segmentation always have been ignored in Uyghur natural language processing and uses the inter-word space as a natural separator simply to obtain the words in the text, is the only word segmentation method so far. But from the Uyghur word’s semantic independence and integrity perspective, can be divided into two categories that single-word structured semantic word (SSSW) and multi-word structured semantic word (MSSW).

Definition 1 (SSSW): Is a Uyghur word also is a no space alphabetic string, complete and independent on its semantics, can be obtains by inter-word spaces based word segmentation (ISWS).

Definition 2 (MSSW): Stable combination of several single-word strings and satisfying the following conditions:

1) Association patterns of two or more words (Commonly is double word or three word structured) and separated by inter-word spaces.
2) Complete and independent on its semantics, and inseparable on its structure.

However, as the Uyghur natural language processing related works constantly goes deeper and wider range of development, ISWS began to expose its potential pitfalls and limitations [1].

In Uyghur Web search, because a MSSW will be split into several fragments that inconsistent with its original meaning, so these word fragments cannot be play to the rule of keywords in text indexing, and also leads to a lot of problems that large scale of word list , big index size, and even low search precision [2].

In the text classification and clustering that the words taken as features, high dimensionality and inter-class cross features is the main factor to restrict the performance of classification and clustering algorithms [3], and the traditional word segmentation method (ISWS) would makes this situation even more serious in Uyghur classification and clustering [4].

Word segmentation also is a bottleneck in machine translation [5], keyword extraction and unknown word processing [6], as well as Uyghur personal name identification (names former surname, separated by a space) and so on [7]. So, the needs of research a kind of automatic word segmentation method and be able to extract the Uyghur words or words associations that stable on its structure grammar, specific and independent on its semantics is increasingly prominent and urgent.

Word segmentation research has a long history in Chinese natural language processing [8], and formed more mature technology and practical segmentation tools [9]. However, the booming popularity of WWW and electronic publications puts forward a series of new
issues on the Chinese automatic segmentation research. Especially the demands of unknown words processing, open environmental adaptability and robustness of word segmentation system have become increasingly prominent. Therefore, more and more scholars have realized that the massive electronic text with extreme ease should be an important resource, and directly obtaining certain applicable knowledge from raw corpus by machine learning method should be an important supplement of automatic word segmentation [10-16].

Our work is closest to that of Maosong S and Sili W, but differs in two important ways. First, we introduced another statistics called dual adjacent entropy (dae), and the liner combinations of difference of t-test (dts), mutual information (mi) and dae as a new measurement to estimate the agglutinative strength of adjacent Uyghur words. Second, we introduced the word association rule (WAR) of Uyghur language features and improved the segmentation accuracy further more.

The basic task of our work is to achieve a Uyghur semantic word segmentation algorithm based on unsupervised learning and non-dictionary strategy, in addition to stemming pretreatment, all statistic information required by the algorithm are automatically derived from raw corpus without manual intervention, and the purpose is to test the algorithm effectiveness in the open environment that completely simulate in practical application.

We have a large corpus collected from the network and formal publications, including the text corpus from Internet (20 classes), all content of the “Xinjiang Daily” in March and April 2008, and eight books (about history, culture, law, politics, economy, etc). In our work, the large corpus is divided into three corpuses and each corpus contains certain percentage of the content in the above large corpus.(1) Uyghur Raw Corpus (URC): contains a total of 9,443,290 Uyghur words and punctuations after segmentation.(2) Uyghur Cooked Corpus(UCU1) for closed test: contains a total of 135708 Uyghur words and punctuations after word segmentation in manual way.(3) Uyghur Cooked Corpus(UCU2) for open test: contains a total of 154411 Uyghur words and punctuations after word segmentation in manual way. All corpuses above are provided by Xinjiang University Key Laboratory of Intelligent Information Processing and also been processed by stem extraction [17].

II. INTER-WORD POSITION JUDGMENT BASED ON STATISTICAL MEASUREMENT

As a semantics specific and independent linguistic unit, internal association degree of a semantic word is relatively close and its relations with the external context are relatively loose [18]. So, it is possible to judge that the words are independent of each other or strongly associated and forms a semantic word according to the association degree between adjacent Uyghur words [19].

The basic approach is to use a statistical measurement S to observe the association degree between adjacent words, if $S > T$ ($T$ for threshold), then keep the “connected” between them, otherwise insert a separator to separate them (because of Uyghur words are separated by spaces, so “)” is used as the separator instead of space), then the adjacent words are “disconnected” and cannot be form a semantic word. For example, $W_1$ is a word string of n words, and inter-word position judgment (“connected” or “disconnected”) based on statistical measurement $S$ as shown in Fig. 1.

![Figure 1. Inter-word position judgment based on statistical measurement](image)

It is easy to calculate the statistical measurement based on the words Bi-gram information be derived from the large scale corpus, such as mutual information, difference of t-test, coupling degree, and so on. On the inter-word position judgment, we can take a kind of basic statistics or some combination of them as the statistical measurement $S$, to estimate the agglutinative strength between adjacent Uyghur words. Of course, uses the combined statistics can be make more accurately judgment than using alone a basic statistics, because it is the comprehensive reflections of association capacity of adjacent words.

In our research, the judgment accuracy of “connected” and “disconnected” between words are taken as an inspection object, constantly adjusting the threshold $T$ and other parameters until the segmentation algorithm is at its best, and the segmentation accuracy $\alpha$ is defined as shown in formula (1):

$$\alpha = \frac{Pos_{con}(x, y) + Pos_{discn}(x, y)}{Pos(x, y)}$$  \hspace{1cm} (1)

In formula (1), $x$ and $y$ are arbitrary adjacent Uyghur words, $Pos_{con}(x, y)$ is stands for the inter-word position numbers of correctly “connected”, $Pos_{discn}(x, y)$ is stands for the inter-word position numbers of correctly “disconnected”, and the $Pos(x, y)$ is stands for the total number of inter-word position in the text.

A. Basic Statistical Measurement: Mutual Information

According to the principles, mutual information $(mi)$ between adjacent Uyghur words $A$ and $B$ is defined as shown in formula (2):

$$mi(A, B) = \log_2 \frac{P(A,B)}{P(A)P(B)}$$  \hspace{1cm} (2)

Among them, $P(A, B)$ is the probability of the adjacent words “A B” appears in the large corpus, $P(A)$ and $P(B)$ are the probability of word A and B appears in the large corpus. $mi(A,B)$ reflects the closeness between adjacent words A and B, if $mi(A, B) \geq 0$, then “A B” is strongly associated; if $mi(A, B) = 0$, then “A B” is weakly associated; If $mi(A, B) < 0$, then “A B” is mutually exclusive. With the
increase of \( m_i(A, B) \), the degree of association is also increased, and if \( m_i(A, B) \) is greater than a given threshold \( T_m \), then we can believe that the “A B” is a semantic word.

We training the Uyghur word Bi-gram model on the raw corpus URC and examine the \( m_i \) judgment distribution about inter-word positions “connected” or “disconnected” on the UCC. Value of \( m_i \) changes in the range between -6.75 and 21.01, when the threshold \( T_m = 4 \) (mean value of \( m_i \) is 3.63 according to URC statistics) the value of \( \alpha \) is up to a maximum of 75.26%. Such as, the judgment of each position on Example 1 is substantially correct, misjudged only one position (As shown in Fig. 2).

**Example 1:**


In this formula, \( p(Y \mid A) \) and \( p(A \mid X) \) are the Bi-gram probability of adjacent words “A Y” and “X A”, \( \sigma^2(P(Y \mid A)) \) and \( \sigma^2(P(A \mid X)) \) are the variance of them respectively. It is can be seen from this formula that, if \( t_{x,y}(A) > 0 \), the association strength of A and with its subsequent Y is greater than with its precursor X, then A should be to break with the X and to be connect with the Y. If \( t_{x,y}(A) < 0 \), the association strength of A and with its precursor X is greater than with its subsequent Y, then A should be to break with the Y and to be connect with the X. If \( t_{x,y}(A) = 0 \), then the association strength between A and X is equal to between A and Y, so it is couldn’t be able to determine if they are should be break or connect.

The t-test is a statistical measurement that based on word statistics and rather than based on the inter-word position, so in order to directly used to the Chinese word segmentation and measuring the association probability of adjacent Chinese words, the t-test formula has been improved and proposed the concept of difference of t-test \((dts)\), and applied to Chinese word segmentation combined with \( m_i \) [10].

According to the definition, the adjacent words “A B” in Uyghur word string “X A B Y” and its \( dts \) value calculation as shown in formula (4):

\[
dts(A, B) = t_{x,y}(A) - t_{x,y}(B) \tag{4}
\]

When \( dts(A, B) > T_m \), the inter-word position of “A B” is more tending to judge for “connected”, and to judge for “disconnected” otherwise. We examined the \( dts \) judgment distribution about inter-words “connected” or “disconnected” still on the UCC.

Changes of \( dts \) in the range between -264.14 and 108.41, and the value of \( \alpha \) is up to a maximum of 78.14% when \( T_m = 0 \). Compared with the \( m_i \), \( dts \) have higher judgment accuracy, but there are some difference with \( m_i \) on the positions judgments in example 1, 2, and 3 (shown in Fig. 5, Fig. 6, and Fig. 7).

**Example 2:**


In this formula, \( p(Y \mid A) \) and \( p(A \mid X) \) are the Bi-gram probability of adjacent words “A Y” and “X A”, \( \sigma^2(P(Y \mid A)) \) and \( \sigma^2(P(A \mid X)) \) are the variance of them respectively. It is can be seen from this formula that, if \( t_{x,y}(A) > 0 \), the association strength of A and with its subsequent Y is greater than with its precursor X, then A should be to break with the X and to be connect with the Y. If \( t_{x,y}(A) < 0 \), the association strength of A and with its precursor X is greater than with its subsequent Y, then A should be to break with the Y and to be connect with the X. If \( t_{x,y}(A) = 0 \), then the association strength between A and X is equal to between A and Y, so it is couldn’t be able to determine if they are should be break or connect.

The t-test is a statistical measurement that based on word statistics and rather than based on the inter-word position, so in order to directly used to the Chinese word segmentation and measuring the association probability of adjacent Chinese words, the t-test formula has been improved and proposed the concept of difference of t-test \((dts)\), and applied to Chinese word segmentation combined with \( m_i \) [10].

According to the definition, the adjacent words “A B” in Uyghur word string “X A B Y” and its \( dts \) value calculation as shown in formula (4):

\[
dts(A, B) = t_{x,y}(A) - t_{x,y}(B) \tag{4}
\]

When \( dts(A, B) > T_m \), the inter-word position of “A B” is more tending to judge for “connected”, and to judge for “disconnected” otherwise. We examined the \( dts \) judgment distribution about inter-words “connected” or “disconnected” still on the UCC.

Changes of \( dts \) in the range between -264.14 and 108.41, and the value of \( \alpha \) is up to a maximum of 78.14% when \( T_m = 0 \). Compared with the \( m_i \), \( dts \) have higher judgment accuracy, but there are some difference with \( m_i \) on the positions judgments in example 1, 2, and 3 (shown in Fig. 5, Fig. 6, and Fig. 7).

**Example 3:**


In this formula, \( p(Y \mid A) \) and \( p(A \mid X) \) are the Bi-gram probability of adjacent words “A Y” and “X A”, \( \sigma^2(P(Y \mid A)) \) and \( \sigma^2(P(A \mid X)) \) are the variance of them respectively. It is can be seen from this formula that, if \( t_{x,y}(A) > 0 \), the association strength of A and with its subsequent Y is greater than with its precursor X, then A should be to break with the X and to be connect with the Y. If \( t_{x,y}(A) < 0 \), the association strength of A and with its precursor X is greater than with its subsequent Y, then A should be to break with the Y and to be connect with the X. If \( t_{x,y}(A) = 0 \), then the association strength between A and X is equal to between A and Y, so it is couldn’t be able to determine if they are should be break or connect.

The t-test is a statistical measurement that based on word statistics and rather than based on the inter-word position, so in order to directly used to the Chinese word segmentation and measuring the association probability of adjacent Chinese words, the t-test formula has been improved and proposed the concept of difference of t-test \((dts)\), and applied to Chinese word segmentation combined with \( m_i \) [10].

According to the definition, the adjacent words “A B” in Uyghur word string “X A B Y” and its \( dts \) value calculation as shown in formula (4):

\[
dts(A, B) = t_{x,y}(A) - t_{x,y}(B) \tag{4}
\]

When \( dts(A, B) > T_m \), the inter-word position of “A B” is more tending to judge for “connected”, and to judge for “disconnected” otherwise. We examined the \( dts \) judgment distribution about inter-words “connected” or “disconnected” still on the UCC.

Changes of \( dts \) in the range between -264.14 and 108.41, and the value of \( \alpha \) is up to a maximum of 78.14% when \( T_m = 0 \). Compared with the \( m_i \), \( dts \) have higher judgment accuracy, but there are some difference with \( m_i \) on the positions judgments in example 1, 2, and 3 (shown in Fig. 5, Fig. 6, and Fig. 7).
C. Basic Statistical Measurement: Dual Adjacent Entropy

As a frequently used language unit, semantic words have a certain degree of circulation in the real text and could be applied in a variety of different context environments, rather than a temporary combination of a special context [21-22]. Therefore, it is possible to estimate the independent linguistic unit possibility of adjacent words according to its contextual environment variability, so as to determine the inter-word position of them is “connected” or “disconnected”.

The contextual variability of adjacent words can be measured by information entropy [23]. To calculate the right and left information entropy of adjacent words “A B”, if they are greater than a given threshold, then can believe that there have varied contextual environment of “A B” and they will be an independent language unit [24], so we can make a judgment that the A and B are “connected”, otherwise “disconnect”.

However, it was found in our segmentation research that, the left and right adjacent entropy were taken as a measurement cannot overall segmentation out the semantic words that three-word structured. For example, making a judgment on the inter-word position of A and B in three-word structured semantic word “A B C”, “A B” may have have varied changes in its left adjacency, but its right adjacency is fixed to the C. According to the definition of information entropy, the right adjacent entropy of “A B” is equal to 0 (minimum), and then the inter-word position between A and B (also between B and C) will be wrongly judged to “disconnect”.

For the above situation, we proposed a kind of statistical measurement called dual adjacent entropy (dae) and very suitable for the research needs of this paper.

Definition 3(dae): For the Uyghur word string “x A B y”, x and y are the contextual right and left adjacency of “A B” each appears in text and constitutes a dual adjacency of sx, sy, then we will have an adjacency set that includes all dual adjacency of “A B” and defines as δsa= {sx, sy}. Supposing the m is stands for the total number of dual adjacency in Ssa,c is stands for the species number (the number of no repeated dual adjacency), n is stands for the frequency of each kind of dual adjacency, then the information entropy (dual adjacent entropy) calculation for dual adjacent set of “A B” as shown in formula(5):

\[ dae(A,B) = -\sum_{i=0}^{n} \frac{n_i}{m} \log \left( \frac{n_i}{m} \right) \]  

It is informed by the formula that, the minimum theoretical value of dae is equal to 0 (when c=1), and the maximum theoretical value is log(m) (when c=m), the greater value of dae(A,B) is indicating that the contextual environment of “A B” is more variable and there is greater possibility of became an independent linguistic unit. If the dae(A,B) is smaller, shows that the independence of “A B” is not strong, and there may be an accidental association between A and B. Therefore, when dae(A,B) > T_dae (T_dae is threshold), the position between A and B is more tending to “connected”, otherwise “disconnected”.

For example, a double-word structured semantic word “A B” appears a total of five times in corpus, and its contextual language environments respectively are “X A B Y”, “Y A B C”, “Z A B X”, “W A B Y” and “V A B X”, dual adjacency set of “A B” is S_{da}= \{<X,Y>, <Y,C>, <Z,X>, <W,Y>, <V,X>\}, m=5, c=5, then dual adjacent entropy of “A B” is:

\[ dae(A,B) = -\frac{1}{5} \log \frac{1}{5} -\frac{1}{5} \log \frac{1}{5} -\frac{1}{5} \log \frac{1}{5} -\frac{1}{5} \log \frac{1}{5} -\frac{1}{5} \log \frac{1}{5} \]  

\[ dae(A,B) = 0.699 \]

We examined the dae judgment distribution about inter-word positions “connected” or “disconnected” still on the UCC1. Changes of dae is in the range between 0.06 and 1.37, and the value of α is up to a maximum of 73.23% when the T_dae = 0.60.

Compared with mi and dts, we have slightly lower segmentation accuracy by dae judgment, and most of the errors occurred on the positions that should be “disconnected”, but it will work well on the unknown words that rarely appeared in the corpus.

For example, Uyghur word A and B are two independent linguistic units that frequently uses in the real text and they also will be adjacent to constitute a specific new word (double-word structured semantic word) “A B”, because of the rarely appearance of “A B” in the large corpus, there will be the maximal of count (A) and count (B), and the minimal of count (A B). In this case, the mi and dts almost making a wrong judgment on the inter-word position, but the judgment of dae is correct, because it is only considered the change diversity of contextual language environment and has little to do with word frequency in dae judgment.

A practical example illustrates this situation in our work, the new word “bird flu” (bird flu) appears in the corpus a total of 17 times, the word “bird flu” appears 2378 times and the word “bird flu” appears 4927 times. Therefore, mi = 3.78, and dts = 0.69, then the inter-word position of “bird flu” is considered “connected” if judged by mi or dts, but the dae = 0.96, then the inter-word position is judged to “connected” by dae. In the following judgment on example 4, mi and dts all are wrong, only the judgment of dae is correct (as shown in Fig. 8, Fig. 9, and Fig. 10).

Example 4:  

create | research | virus | bird flu | scientist  

Figure 8. dae segmentation results on the Example 4

However, it is the mi, dts or dae, because they all are the basic statistical measurement and only considered one
aspects of statistical information, so there are some inevitable limitations if uses only one kind of them to measuring inter-word position.

\[
\begin{align*}
\text{ds} & = -264.14 \text{ to } 108.41, \\
\text{mi} & = -6.75 \text{ to } 21.01, \\
\text{dae} & = -6.51 \text{ to } 3.63, \\
\end{align*}
\]

The \( \lambda = 0.35 \) and \( \gamma = 0.30 \). The variation range of \( \text{ds} \) on UCC is between -11.5 and 6.9, and the \( \alpha \) is up to a maximum of 84.31% (T_{\text{tud}} = 0), respectively increased by 6.17%, 9.05% and 11.08% compared with separately using of \( ds, mi \) and \( dae \).

In our segmentation experiment on the example 4, \( \text{ds}(\text{ئۇيغۇرچە ئۇئەرپەس}, \text{ئۇئەرپەس}) = 1.67, \text{mi}(\text{ئۇيغۇرچە ئۇئەرپەس}) = 3.87, \) and \( \text{dae}(\text{ئۇيغۇرچە ئۇئەرپەس}) = 0.96, \) then there have wrong judgments (“disconnected”) made by \( ds, mi, \) and \( dae \) judgment is correct. After normalization, \( \text{dmd} = 0.38 > T_{\text{tud}}, \) and then the inter-word position (ئۇيغۇرچە ئۇئەرپەس) is correctly judged to “connected” after uses the \( \text{dmd} \) measurement.

For the inter-word position (ئۇيغۇرچە ئۇئەرپەس) also on the example 4, \( ds(\text{ئۇيغۇرچە ئۇئەرپەس}, \text{ئۇئەرپەس}) = 5.46, \text{mi}(\text{ئۇيغۇرچە ئۇئەرپەس}) = 4.11, \) and \( \text{dae}(\text{ئۇيغۇرچە ئۇئەرپەس}) = 0.63, \) then there have wrong judgment (“disconnected”) made by \( ds, mi, \) and \( dae \) judgment is correct. After normalization, \( \text{dmd} = 0.07 > T_{\text{tud}}, \) and then the inter-word position (ئۇيغۇرچە ئۇئەرپەس) is correctly judged to “connected” after uses the \( \text{dmd} \) measurement (as shown in Fig. 11).

It is thus clear that, \( \text{dmd} \) judgment is consistent with \( ds, mi, \) and \( dae \) when they have agreements on their judgment, and also there is a certain degree of complementarity when their judgment is not consistent.

III. LINGUISTIC FEATURES AND WORD ASSOCIATION RULE

In our work, using the combined statistical measurement \( \text{dmd} \) to estimate the inter-word positions and the segmentation accuracy reached 84.31%, but there is still some distance with the accuracy of the ideal. Therefore, we look for some information from the Uyghur language itself that could be helpful for inter-word position judgment, and found the following linguistic features different from Chinese.

nouns or verbs. Therefore, adjacent words like “N+ADJ” or “V+ADJ” relations never to be constitute a semantic word.

In our work, we have statistics the number of independent words and adjectives and their frequency on the raw corpus URC. There are 592 independent words with total frequency of 1634860, accounted for 17.3% of the corpus. Adjectives are 988 with total frequency of 1170968, accounted for 12.4% of the corpus. We also went to observe the inter-word positions that wrongly judged by $dmd$ on the UCC1, found that 3.96% of 15.69% errors just occurred on the independent words and adjectives, and all the “disconnected” be wrongly judged to “connected”.

According to the above situations and also in accordance with the Feature 1 and Feature 2, the rules uses for inter-word position judgment defined as follow.

**Definition 4** (word association rule: WAR): For the adjacent Uyghur words “A B”, IW stands for the independent words, $\text{Pos}(A, \ B)$ stands for the inter-word position of “A B”, ADJ stands for the adjectives, then:

\[
\begin{align*}
\text{If } A \in \{\text{IW}\} \text{ or } B \in \{\text{IW}\} & \text{ then } \\
\text{Pos}(A, B) & \leftarrow \text{“disconnected”} \\
\text{else if } B \in \{\text{ADJ}\} & \text{ then } \\
\text{Pos}(A, B) & \leftarrow \text{“disconnected”} \\
\text{else} & \text{ then } \\
\text{Pos}(A, B) & \leftarrow \text{“connected” or “disconnected”}
\end{align*}
\]

Therefore, we have established two kind of word list that separately includes independent words and adjectives; the purpose is to judge theinter-word positions whether like feature 1 or feature 2 based on the word list. In this way, on the one hand to reduce the $dmd$ computations for considerable part of inter-word positions; on the other hand, to judge inter-word positions use the known rule of linguistic features, this will very helpful to improve the accuracy.

**IV. SEMANTIC WORD SEGMENTATION ALGORITHM BASED ON DMD AND WAR**

After the determination of combined statistical measurement $dmd$ and the WAR in accordance with linguistic features, the Uyghur semantic word segmentation algorithm is as follows:

**Step 1:** Do stem extraction for the text will be processed, and all the punctuations replaces with predefined separator “|”.

**Step 2:** For the inter-word position of any adjacent adjectives “A B”, calculate the $dmd$ values of all inter-word positions in the text to be processed under the following conditions:

\[
\begin{align*}
\text{if } A & \in \{\text{IW}\} \text{ or } B \in \{\text{IW}\} \text{ or } B \in \{\text{ADJ}\} \text{ then } \\
\text{dmd}(A, B) & = -1 \text{ (less than the } T_{\text{max}}) \\
\text{else} & \text{ then } \\
\text{dmd}(A, B) & = \text{dts}(A, B) + \alpha \times m^i(A, B) + \gamma \times \text{dae}(A, B)
\end{align*}
\]

**Step 3:** In turn, to extract the current inter-word position, perform step 4 until the end.

**Step 4:**

\[
\begin{align*}
\text{If } \text{dmd} > T_{\text{max}} & \text{ then } \\
\text{Pos}(A, B) & \leftarrow \text{“connected”} \\
\text{else} & \text{ then } \\
\text{Pos}(A, B) & \leftarrow \text{“disconnected”}
\end{align*}
\]

The open test and closed test experimental results under different segmentation strategy are summarized as shown in Table II and Table III.

**TABLE I. WORDS AND INTER-WORD POSITIONS IN URC1 AND URC2**

<table>
<thead>
<tr>
<th>Test set</th>
<th>Inter-word positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open test set (URC1)</td>
<td>120948</td>
</tr>
<tr>
<td>Closed test set (URC2)</td>
<td>137757</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE II. CLOSED TEST EXPERIMENTAL RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmentation strategy</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>dts</td>
</tr>
<tr>
<td>mi</td>
</tr>
<tr>
<td>dae</td>
</tr>
<tr>
<td>dmd</td>
</tr>
<tr>
<td>dmd +WAR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE III. OPEN TEST EXPERIMENTAL RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmentation strategy</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>dts</td>
</tr>
<tr>
<td>mi</td>
</tr>
<tr>
<td>dae</td>
</tr>
<tr>
<td>dmd</td>
</tr>
<tr>
<td>dmd +WAR</td>
</tr>
</tbody>
</table>

It is can be seen from the experimental results that, the algorithm performance has not decreased in the open test. In addition, the accuracy of $dmd$ segmentation is invariably higher than that dts, mi, and dae, indicating that the proposed measurement $dmd$ and its parameters determination are valid. Especially, $dmd$ judgment will be more accurate with the assist of WAR.

**B. Problems and Future Work**

In the experiment, we also found the following problems affecting the accuracy of Bi-gram statistical
information to a certain extent, and leads wrongly judgment on the inter-word position.

(1) Limitations of the stem extraction tool. Because of the existing stemming algorithm is more dependence on the pre-built stem library, so it is hard to work on the stem extraction for the unknown words.

(2) Spelling errors hard to avoid. There are more similar letters in Uyghur alphabet, and in addition the custom difference of each place people pronunciation (dialect), the correct spelling is not always an easy task in Uyghur language, such as even many Uyghur people cannot exactly spell the word: "ئۇزونىرىستيمي" (university). Therefore, spelling error is a main factor influencing the pattern matching and word frequency statistics.

(3) Non-standard abbreviations. Now that, the authorities have set the use the standard specification for the new words from foreign languages, but there are still arbitrariness of non-standard abbreviations on the internet. For example, the standard terminology "คอมพิวเตอร์" (computer) is abbreviated as "كوم" or "өй"; "ایمیل" (E-mail) is abbreviated as "ئىمیل" or "ئېمیل" etc. This will also influencing the pattern matching and word frequency statistics.

For the limitations of stem extraction, spelling error is also a major factor in addition to the shortcomings of the algorithm itself, and the existing Uyghur NLP methods and tools can not be effective on the fully automatic error detection and correction for batch processing to large text [25], and about problems (3), there have not any related research report up to now.

Obviously, whether it is the imperfections of text pretreatment algorithm or defects of the original text, all this situations will affects the term extraction accuracy and also affects the index quality and search precision. Therefore, as far as possible to eliminate these negative factors before semantic word extraction and to obtain the higher extraction accuracy on the normative text corpus is completely possible, this will be the research emphasis in our future work.

VI. CONCLUSION

This paper discusses the traditional concept of Uyghur word segmentation, explores the limitations and problems causes by the commonly used method, and puts forward the idea of Uyghur semantic word segmentation and a kind of automatic segmentation algorithm for the first time. In this algorithm, no longer takes the inter-word space as a natural delimiter, but to estimate the agglutinative strength between adjacent Uyghur words and identifies the segmentation positions based on the strategy of combined statistics and rules.

Motivations and reasons of the strategy to use are in the following two aspects: One of the main reasons is that there never have word segmentation idea and the dictionary for word segmentation in Uyghur language. The second reason is that the many successful stories in Chinese word segmentation have proved the effectiveness of this method. Therefore, We have directly obtained words Bi-gram statistics from large-scale raw corpus, introduced a new statistics that called dual adjacent entropy(dua), and take the liner combinations of mutual information, difference of t-test and dual adjacent entropy as a combined measurement (dmd) to estimate the inter-word positions which is connected or disconnected. When calculating the dmd, we also introduced the language features and improved the segmentation accuracy with the aid of Uyghur word combination rules.

Proposed algorithm shows the higher segmentation accuracy and robustness on the large-scale open test, but there is a lot of works have to continue in the mechanisms and methods, as well as the segmentation system and its verification.

ACKNOWLEDGMENT

This work has been supported by the National Natural Science Foundation of China (61063022, 61262062, 61163033, and 61262063), Scientific Research Program of the Higher Education Institution of Xinjiang (XJEDU2012111), and Program for New Century Excellent Talents in University (NCET-10-0969).

REFERENCES


Turdi Thohti born in 1975, received B. E. in 1999 from Nanjing University of China, received M. E. in 2009 from Beijing University of technology of China. He is an associate professor and Graduate Supervisor in the School of Information Science and Engineering, Xinjiang University. He has published more than 30 papers on international journals, national journals, and conferences in recent years. His research interests include natural language processing, information retrieval, web mining and content security. He is a PhD candidate in Xinjiang University, Senior of CCF, member of IEEE CS and CIPSC.

Winira Musajan born in 1960, received B. E. in 1983 from Xinjiang University of China. She is a professor and Graduate Supervisor of School of Information Science and Engineering, Xinjiang University. Her main research interests in natural language processing and multilingual information retrieval.

Askar Hamdula born in 1972, received B. E. in 1996, M. E. in 1999, and Ph. D. in 2003, all in Information Science and Engineering, from University of Electronic Science and Technology of China. In 2010, he was a visiting scholar at Center for Signal and Image Processing, Georgia Institute of Technology, GA, USA, tutored by Professor Biing-Hwang (Fred) Juang. Currently, he is a professor in the School of Software Engineering, Xinjiang University. He has published more than 80 technical papers on speech synthesis, natural language processing and image processing. He is an affiliate member of IEEE, member of phonetics branch committee of Association of Chinese Linguistics.