Typing Behaviour Gives Chinese Word Segmentation New Perspective

Dakui Zhang and Yu Mao

***Abstract*—Users provide a lot of undiscovered information when they are working with electronic products, such as computer. Many natural annotations are left when users type Chinese materials with input method. We call them natural typing annotations. These annotations have intimate connection with Chinese word segmentation. In this paper we collect corpus with natural inputting annotations and analyze it in various respects. From the corpus, user's typing patterns are extracted and classification model is built to identify different patterns. Experiments show that natural inputting annotations have promising potential in overcoming the drawbacks of existing word segmentation approaches.**

***Index Terms*—Natural inputting annotations, Chinese word segmentation, user's typing patterns, classification model, voting mechanism.**

I. Introduction

Many natural annotations for segmentation are left when users type in text materials, such as space characters in English. Traditional Chinese text materials do not include explicit delimiters among words. Therefore, word segmentation is a necessary initial step for Chinese language processing. Indeed, invisibly natural annotations exist during the process of typing Chinese text materials.

According to our own experience in using Chinese input methods, users need to confirm their inputting content with space key, number key or enter key frequently. However, unlike space in English, this information is not visually recorded in Chinese text. For example, when users type in '我在写论文(I am writing thesis)', one of the probable sequences is '我<SPACE>在<SPACE>写<SPACE>论文<SPACE>'. <SPACE> stands for space key that user used to confirm inputting content. These visual confirmation tags recorded between words are one kind of natural annotations. We call them natural typing annotations [1].

Natural typing annotations provide a new perspective to rethink about the segmentation of Chinese words. This paper is focused on collecting the corpus with natural typing annotations and exploring the relationship between user's typing pattern and word segmentation. We analyze the collected corpus in various respects and extract three user's typing patterns. Classification model is built to identify different patterns and the method to find proper corpus is proposed. Experimental results show that natural typing annotations have promising potential in overcoming the drawbacks of existing word segmentation approaches.

II. Collection and Analysis of Corpus

When users typing in Chinese characters with Chinese input methods, natural annotations used for confirming the input content would be left during the process. These natural annotations could reflect the behavioral and psychological habits of users, and also have close relationship with word segmentation.

A. Corpus

When you submit your final version, after your paper has been accepted, prepare it in two-column format, including figures and tables.

So far, there are no open corpora with the information about confirming the inputting content during the typing process. We need to collect related data independently.

Chinese input methods are the typical way to produce Chinese text. The typical way to type in Chinese characters is in a sequential manner [2]. Take Chinese Pinyin input methods for example. Assume users wants to type in Chinese word '背景 (background)'. First, they mentally generate and type in corresponding Pinyin ' beijing '. Then, a Chinese Pinyin input method displays a list of Chinese words which share that Pinyin. Finally, users choose the target word from candidates with the space key or the corresponding numeric key. So are the input processes of Wubi and other Chinese input methods. The traditional Chinese text does not include the user's confirming information. In order to make those information visible, we put a '|' after the content users confirmed.

We consider what between two punctuation marks as a sentence $S = c_1 c_2 ... c_N$ (c_i stands for a Chinese character). After input by the user, the sentence with confirming information is segmented into $\pi(S) = c_1 ... c_{i_1 - 1} \mid c_{i_1} ... c_{i_2 - 1} \mid ... \mid c_{n_1} ... c_N \mid$. '|' is the natural typing annotation left during the process. What between two '|'s is a segment. Then the segmentation, $\pi(S) = segment_1 \mid segment_2 \mid ... \mid segment_M \mid$ ($M \le N$), is considered as corpus with natural typing annotations.

For comparison, test text with ambiguous meaning, named entities or promiscuous words is chosen for the experiment. Participants should manually type in the test text and leave natural typing annotations in the results. Three examples from test text are listed in Fig. 1.

Manuscript received December 26, 2016; revised March 1, 2017. This work was supported by the National Strategic Basic Research Program ("973" Program) of the Ministry of Science and Technology of China (No. 2012CB720702 and 2013CB329303).

The authors are with the Beijing Institute of Technology, Beijing 100081, China (e-mail: dakuiz@163.com; maoyu1998@qq.com).

DOI: 10.7763/IJET.2017.V9.1023

Finally, the typing corpus of 384 participants was collected after experiment A.

B. Analysis of Corpus

In order to analyze the whole situation of the 384 participants' input, we sort out 66,232 segments in the corpus, and get 883 non-redundant ones. Then the frequencies of the different segments also are counted. The results show that the segments with frequency more than 10 are individual characters, or simple phrases and expressions with no more than 4 Chinese characters. This means the behavioral and psychological habit of using simple word, phrase or expression as a segment actually exists. This habit also meets the principle of behavioral economics. People consciously avoid the mistakes that might be brought by inputting long material one time. Besides, people seldom put the words with no logical meaning in one segment. Taking '主人公严守一把手机给扔了。(The leading character Yan Shouyi has thrown his cellphone away.)' for example, when participants input '给扔了(have thrown)', they choose to type in the material as '|给|扔|了|', '|给|扔了|' or '|给扔了|'. No one types in the material as '|给扔|了|', because '|给扔|' has no logical meaning in Chinese. So the constitution of segment can reflect the language logic of the participants.

主人公严守一把手机给扔了。

袁帅的英文名字叫卡萨诺瓦，他一点都没有元帅的英武之气，伊丽莎白就不喜欢他。

华硕和三星的电脑都不错，不过评价在三星级以上的这几款电脑，我感觉华硕的更划算一些。

Fig. 1. Three examples from test text.

TABLE I: RELATIVE FREQUENCIES OF SEGMENT LENGTH FROM THREE SEGMENTATION RESULTS

Length(seg)	Relative frequency		
	User's input	*Gold standard*	*ICTCLAS*
1	40.51%	48.57%	55.31%
2	43.36%	47.14%	42.09%
3	9.55%	3.33%	2.70%
4	3.06%	0.95%	0.90%
5	1.09%	0	0
6	0.49%	0	0
7	0.58%	0	0
8	0.38%	0	0
9	0.25%	0	0
10	0.18%	0	0
11	0.24%	0	0
12	0.07%	0	0
13	0	0	0
14	0.07%	0	0
15	0	0	0
16	0.09%	0	0
17	0.04%	0	0
18	0.04%	0	0

$Length(seg)$ is used to express the length of a segment. Then, a frequency distribution of different $Length(seg)$ can be sorted out from the input results of the 384 participants. Moreover, we segment the test text with gold standard and ICTCLAS segmenter [3] separately. The same statistics are done with the two results to get frequency distributions. At last, the two frequency distributions multiplied by 384 separately are two final results, which can be compared with inputting results of the participants. As shown in Table I.

III. IDENTIFICATION OF HIGH QUALITY CORPUS

In this section, user's typing patterns are extracted at the sentence level. Classification model is built to identify sentences with different patterns.

A. Three User's Typing Pattern

Different users use different typing patterns to input Chinese. S_1 = '不过评价在三星级以上的这几款电脑 (However, these several computers are all with assessments more than 3 stars)' is taken as an example to explain the different situations. Just as what is shown in the following, $\pi_{gold}(S_1)$ is the gold standard word segmentation of S_1, others are the results of different users.

$\pi_{gold}(S_1)$ = '|不过|评价|在|三星级|以上|的|这|几|款|电脑|'

$\pi_1(S_1)$ = '|不过|评价|在|三星级|以上|的|**这几款**|电脑|'

$\pi_2(S_1)$ = '|不过|评价|在|**三|星|级**|以上|的|**这几**|款|电脑|'

$\pi_3(S_1)$ = '|**不过评价**|在|**三星级以上**|的|**这几款电脑**|'

$\pi_4(S_1)$ = '|**不过评价在三星级以上的**|**这几款电脑**|'

$\pi_5(S_1)$ = '|**不|过**|评价|在|**三|星|级|以|上**|的|这|几|款|**电|脑**|'

With observation to the corpus with natural typing annotations collected from the users, three typical phenomena of typing in Chinese words are found. The first one is segregation, which means that the characters belong to one segment in the result of gold standard are separated into different segments. For example, '电脑(computer)' is actually one word, but in $\pi_5(S_1)$, this word is separated into two segments as '|电|脑|'. The second one is adhesion. It means two or more adjacent individual words in the result of gold standard are glued together into one segment. For example, '这几款(this several)' , three independent words according to gold standard, are put together in one segment in $\pi_1(S_1)$ as '|这几款|'. The third one is moderation, which means that the words from the result of gold standard are put into different segments properly.

At sentence leave, there are three user's typing patterns in the corpus collected from the users. The first user's typing pattern is the gold segmentation, which means that the natural typing annotations left by users are same as in the result of gold standard. No segregation and adhesion appear in a sentence, just as $\pi_{gold}(S_1)$. The second user's typing pattern is that few (no more than 30% of gold standard, 30% is an empirical value) segregation or adhesion phenomena exist in a sentence, just like $\pi_1(S_1)$ and $\pi_2(S_1)$. The third user's typing pattern is many (more than 30%) segregation or adhesion phenomena appear in a sentence, just like $\pi_3(S_1)$, $\pi_4(S_1)$ and $\pi_5(S_1)$.

For sentences following the first and second patterns, they are very close to each other and will expert positive effects on word segmentation. We call such sentences high quality corpus. Sentences in the third pattern, which differ greater with the result of gold standard, are the low quality corpus. Generally speaking, user's typing habits would not be changed frequently. So, who are accustomed to inputting high quality corpus are called users with high quality typing habits, and others are called users with low quality typing habits. Obviously, high quality corpus plays a positive role in doing word segmentation. More effective method for identifying high quality corpus needs to be found.

B. Identification of High Quality Corpus

Identification of high quality corpus can be handled as classification problem. Effective and logical feature should be found to identify the corpus with natural typing annotations. The feature of high quality corpus is that the whole number of the Chinese characters in one sentence and the number of the segments in one sentence are maintained in a rational level, which means that neither the large number of segments with only one character nor the number of segments with a lot of characters would appear in a high quality sentence. Five simple but distinguishing features can be used to classify sentences.

Len is the length of the sentence.

SegNum(SN) stands for the number of the segments in the sentence.

These two features can be used to figure out whether the percentage of character number of the sentence and the segment number of the sentence is in a proper range.

SingleSegNum(SSN) stands for the number of the segments with only one character in the sentence.

MaxConSingleSegNum(MCSSN) is the maximum number of continuous segments with only one character.

MaxSegLen(MSL) means the length of segment with most characters.

These three features can be used to figure out whether there are many segregation and adhesion phenomena in the sentence.

Take $\pi_5(S_1)$ for example, Len=16, SN=15, SSN=14, MCSSN=12, MSL=2.

All the sentences can be labelled as **Good**, **Normal** or **Bad**, which correspond to the first, second and third user's typing pattern separately.

Support Vector Machine(SVM) [4] is chosen as the classifier for the experiment. SVM is a machine learning algorithm based on statistical learning theory. It has a strong generalization ability. Even though the number of samples is small, a good statistical regularity can be efficiently got.

Experiment B shows that features mentioned in this paper could distinguish high quality corpus and low quality corpus effectively. But the identification ability between **Good**, **Normal** is relatively poor.

C. Ranking Mechanism for High Quality Corpus

There are different forms for the high quality corpus got from classification models. For example, $\pi_{gold}(S_1)$, $\pi_1(S_1)$ and $\pi_2(S_1)$ in section III.A are all high quality corpus. Ranking mechanism is introduced to learn which one is the most approved one among all these forms of high quality corpus.

Assume k stands for the number of non-redundant segmentations $\pi_1(S)$, $\pi_2(S)$,…, $\pi_k(S)$ for sentence S, then $count(\pi_i(S))$ calculates how many users input S with segmentation $\pi_i(S)$. Then most approved inputting segmentation by the users can be expressed as (1).

$$\pi_{approved}(S) = \arg\max_{\pi_1(S)}^{\pi_k(S)} count(\pi_i(S)) \quad (1)$$

It means that $\pi_i(S)$ gets one more vote when it is repeatedly input by users. The one with most votes is the most-approved-segmentation. Taking the sentence S_1 in section III.A for example, $\pi_1(S_1)$ is repeatedly input ten times, which is more than any other segmentation. Therefore, we consider it as the most-approved-segmentation for the users. But actually this is not a right word segmentation result, because '|这几款|', taken as one segment in $\pi_1(S_1)$, is actually three individual segments in $\pi_{gold}(S_1)$. On the other hand, such logical chunks are more valuable in some application scenarios, such as machine translation, than results of gold standard.

With identification tools for high quality corpus and ranking mechanism for repeatedly inputting sentences, the typing habits of one user can be estimated generally. Assuming there is a window, the ratio of high quality corpus in this window can reflect the result. Taking 20 sentences as a window for a user's input, if the ratio of high quality corpus is higher than 85% in this window, this user is a user with better typing habits. The size of the window and the ration of high quality corpus could be adjusted according to the specific corpus and different identification accuracy requirement.

Among users with the better input habits, user ranking can be further conducted through ranking mechanism. In a user's typing corpus, the more most-approved-segmentations exist, the higher this user is ranked.

In experiment C, we get a word segmentation result of the test text through most-approved-segmentations. And the users who are ranked at top 3 positions are found. Experimental results show that Collective Intelligence gives a new perspective on solving word segmentation.

IV. Experiments

A. Collection of Participants' Typing Corpus

Software is developed to collect the natural typing annotations left during the inputting process. However, another key point of this experiment is how to gather the volunteer participants.

A public notice is posted on the Internet to gather volunteer participants. There are some requirements for the volunteers. First of all, volunteers spend more than 3 hours with computer every day. Second, volunteers use Social Networking Services frequently. Third, they should be over eighteen-year old. Their input habits and Chinese input methods they used to employing are not learned before the experiment. Finally, 384 participants' typing results are valid. Altogether, there are 1089 sentences, 5517 segments, among which 635 are non-redundant.

According to the survey after the experiment, these 384 participants spend at least more than 3 hours with computer every day and they use Chinese as their input language in their work and life. The youngest participant is 18 and the oldest one is 65. The male to female ratio of participants is close to 1:1. There are senior high school graduates, undergraduates and graduate students. Among these 384 participants, 381 used to employ pinyin input method, 3 used to employ Wubi input method. 2009 China Desktop Software Development Research Report [5] issued by iResearch says that in 2009, 53.2% users choose to employ Sogou Pinyin input method, 16.9% choose to employ Wubi input methods, 10.3% Microsoft Pinyin input method and 5.7% Google Pinyin input method. According to this report, the participants chosen for our experiment are in the normal range which had no huge difference compared with the majority users in China.

B. Classification of Corpus

Package of libSVM [6] is used in this experiment. Radial basis function is adopted as kernel function, gamma value equals to 1/num_features and cost value is 1.

10-fold cross validation is used to validate the results. The 1,089 sentences are partitioned into ten parts randomly. Nine of ten are chosen as training set, the left one is testing set. It is conducted ten times and every part should be testing set once. Classification accuracy of the experiment is listed as the second column in the Table II below.

TABLE II: 10-FOLD CROSS VALIDATION RESULTS

Num	3-Classification	2-Classification
1	81.65%	96.33%
2	79.69%	97.22%
3	86.24%	97.25%
4	78.90%	97.25%
5	76.15%	89.91%
6	80.73%	98.17%
7	77.98%	94.50%
8	81.92%	94.59%
9	86.36%	94.55%
10	86.79%	98.11%
Average	81.64%	95.79%

It is found that the classification errors occur mainly between the Label Good and the Label Normal. Now that both of them are the high quality corpus, the Label Normal can combine with the Label Good. If there are only two labels: Good and Bad, accuracy is shown as third column in Table II.

Besides the test text that every participant has to input, some participants are also asked to provide corpus they input in daily time with natural typing annotations. 50 sentences are chosen to be included in the set for verifying the generalization ability of the model. According to the result of the classification, with arbitrary typing, the average accuracy of the 3-classification is 78.00%, and the 2-classification is 90.00%, which means that the model we developed has effective identification ability in distinguishing high quality corpus and low quality corpus.

C. Ranking to Find High Quality Corpus

According to ranking mechanism, every sentence S in the test text has the most-approved-segmentation $\pi_{approved}(S)$ from the high quality corpus. A word segmentation result of test text can be got from the most-approved-segmentations. It is considered as word segmentation conducted by user's input.

We get another word segmentation result of the test text from ICTCLAS segmenter. Gold word segmentation of the test text is manually done following the standard of MSA corpus in bakeoff 2005. In evaluating segmentation accuracy, we used three measures: precision, recall and balanced F-score. Precision p is defined as the number of correctly segmented words divided by the total number of words in the automatically segmented corpus. Recall r is defined as the number of correctly segmented words divided by the total number of words in the gold word segmentation. F-score f is defined as follows:

$$f = \frac{p \times r \times 2}{p + r} \tag{2}$$

The results are tabulated in Table III.

TABLE III: WORD SEGMENTATION RESULTS

Word segmentation from	***p***	***r***	***f***
User's input	86.19%	74.29%	79.80%
ICTCLAS	85.65%	90.95%	88.22%

We find two main reasons for the errors in word segmentation result got from user's input. The first one is that pronoun is adhered with the word or phrase fore-and-aft as one segment. For example, '大家好(hello)', '我叫(my name is)', '这就是(this is)', '让自己(let myself)' are all taken as one segment. The second one is that auxiliary word is adhered with the word or phrase fore-and-aft as one segment. For example, '扔了(threw)', '写了(wrote)', '大的(big)', '小的(small)'. These errors can be handled with simple rules. We can easily separate pronouns or auxiliary words from the original segments. At the same time, the three-label classification can be adopted, and more weight can be given to Label Good during the ranking mechanism. After these two steps, amended results are shown in Table IV.

TABLE IV: AMENDED WORD SEGMENTATION RESULTS

Word segmentation from	*p*	*r*	*f*
User's input	94.42%	88.57%	91.40%
ICTCLAS	85.65%	90.95%	88.22%

Besides the evaluation in overall situations, details are also evaluated. First of all, different persons' names are arranged in the test text. Identification ability for named entity(NE) is listed in the Table V.

Secondly, In sentence '主人公严守一把手机给扔了。(The leading character Yan Shouyi has thrown his cellphone away.)', '严守一(Yan Shouyi.)' is a person's name. At the same time, '严守(strictly observe)' and '一把(手)(a/leader)', '一把（手）(a/leader)' and '手机(cellphone)' also have overlapped ambiguity. The result we get from the ICTCLAS is '主人公|严守|一把手|机|给|扔|了|。'. The result we get from user's input is '主人公|严守一|把|手机|给|扔了|。'. For a long period, out-of-vocabulary(OOV) words and segmentation ambiguity are two main influencing factors for the accuracy of segmentation [7]. It can be seen that the natural typing annotations play positive role in solving the problems of traditional word segmentation algorithms.

TABLE V: NAMED ENTITY RECALL RESULTS

Word segmentation from	The num of NEs appear in test text	The num of NEs segmented correctly	*r*
User's input	8	8	100.00%
ICTCLAS		5	62.50%

Through ranking mechanism, top 3 users with better input habits are also listed out. Table VI show how close their corpus with natural typing annotations can be to gold standard.

TABLE VI: WORD SEGMENTATION RESULTS FROM TOP 3 USERS

Word segmentation from	*p*	*r*	*f*
User#254	87.57%	77.14%	82.02%
User#011	86.15%	80.00%	82.96%
User#128	83.33%	73.81%	78.28%
ICTCLAS	85.65%	90.95%	88.22%

Adhesion phenomena also exist in their input. Same strategies are used on their corpus. The amended results are tabulated in Table VII.

TABLE VII: AMENDED WORD SEGMENTATION RESULTS FROM TOP 3 USERS

Word segmentation from	*p*	*r*	*f*
User#254	92.82%	90.19%	91.49%
User#011	91.50%	88.29%	89.87%
User#128	90.38%	87.33%	88.83%
ICTCLAS	85.65%	90.95%	88.22%

All the results show that the corpus left by users with better input habits is very close to the gold word segmentation. With some simple processing, like adding or deleting some separators ('|'), this corpus can be used as training corpus for word segmentation.

V. CONCLUSION AND EXPECTATION

The natural typing annotations left by users during their typing process are valuable information. They exist always, but are neglected by us for a long period. In this paper, we collect the corpus with natural typing annotations, and then analyze the features of user's input. After extracting three user's typing patterns, we build up classification model to identify different patterns. The experiments show that the natural typing annotations play positive role in improving the existing word segmentation algorithms.

For future work, we will continue to collect more corpus with natural annotations, then develop word segmentation algorithm suitable for corpus with natural typing annotations, and explore more information left by users during their typing process.

REFERENCES

[1] D. K. Zhang, Y. Mao, Y. Liu *et al*., *The Discovery of Natural Typing Annotations: User-produced Potential Chinese Word Delimiters*, In ACL-IJCNLP, 2015.

[2] J. T. Wang, S. M. Zhai, and H. Su, *Chinese Input with Keyboard and Eye-Tracking: An Anatomical Study*, 2001.

[3] H. P. Zhang, H. K. Yu, D. Y. Xiong, and Q. Liu, *HHMM-based Chinese Lexical Analyzer ICTCLAS*, 2003.

[4] H. Yamada and Y. J. Matsumoto, *Statistical Dependency Analysis with Support Vector Machines*, 2003.

[5] (2009). China Desktop Software Development Research Report. [Online]. Available: http://report.iresearch.cn/1290.html.

[6] C. C. Chang and C. J. Lin, *LIBSVM: A Library for Support Vector Machines*, In TIST, 2011.

[7] C. N. Huang and H. Zhao, "Chinese word segmentation: A decade review," *Chinese Information Processing*, 2003.

Dakui Zhang received M.S. degree in computer software and theory from the Henan Polytechnic University in 2011. He is now Ph.D. candidate at Beijing Institute of Technology, China. His research interests include natural language processing, deep learning, data mining.

Yu Mao is now Ph.D. candidate of computer software and theory at Beijing Institute of Technology, China. His research interests include natural language processing, machine learning.