# The Discovery of Natural Typing Annotations: User-produced Potential Chinese Word Delimiters

**Dakui Zhang[1], Yu Mao[1], Yang Liu[1], Hanshi Wang[2], Chuyuan Wei[1], Shiping Tang[1]**

[1]Beijing Institute of Technology, Beijing, China
[2]Capital Normal University, Beijing, China
{sbirdge,maoyubit,yan9liu,necrostone,weichuyuan}@gmail.com, simontangbit@bit.edu.cn

## Abstract

Human labeled corpus is indispensable for the training of supervised word segmenters. However, it is time-consuming and labor-intensive to label corpus manually. During the process of typing Chinese text by Pingyin, people usually need to type "space" or numeric keys to choose the words due to homophones, which can be viewed as a cue for segmentation. We argue that such a process can be used to build a labeled corpus in a more natural way. Thus, in this paper, we investigate Natural Typing Annotations (NTAs) that are potential word delimiters produced by users while typing Chinese. A detailed analysis on over three hundred user-produced texts containing NTAs reveals that high-quality NTAs mostly agree with gold segmentation and, consequently, can be used for improving the performance of supervised word segmentation model in out-of-domain. Experiments show that a classification model combined with a voting mechanism can reliably identify the high-quality NTAs texts that are more readily available labeled corpus. Furthermore, the NTAs might be particularly useful to deal with out-of-vocabulary (OOV) words such as proper names and neo-logisms.

## 1 Introduction

Unlike English text in which sentences are sequences of words delimited by white spaces, in Chinese text, sentences are usually represented and stored as strings of Chinese characters without similar natural delimiters. To find the basic language units, i.e. words, segmentation is a necessary initial step for Chinese language processing.

Currently most of state-of-the-art methods for Chinese word segmentation (CWS) are based on supervised learning, which depend on large scale annotated corpus. These supervised methods obtain high accuracies on newswire (Xue and Shen, 2003; Zhang and Clark, 2007; Jiang et al., 2009; Zhao et al., 2010; Sun and Xu, 2011). However, manually annotated training data mostly come from the news domain, and the performance can drop severely when the test data shift from newswire to blogs, computer forums, and Internet literature (Liu and Zhang, 2012;).Supervised approaches often have a high requirement on the quality and quantity of annotated corpus, which is always not easy to build. As a result, many previous methods utilize the information of free data which contain limited but useful segmentation information over the Internet, including large-scale unlabeled data, domain-specific lexicons and semi-annotated web pages such as Wikipedia. There has been work on making use of both unlabeled data (Li and Sun, 2009; Sun and Xu, 2011; Wang et al., 2011; Qiu et al., 2014) and Wikipedia (Jiang et al., 2013; Liu et al., 2014;) to improve segmentation. But none of them notice the segmentation information produced by users while typing Chinese.

Chinese is unique due to its logographic writing system. Chinese users cannot directly type in Chinese words using a QWERTY keyboard. Input methods have been proposed to assist users to type in Chinese words (Chen, 1997). Substantial information has been produced, but not recorded and stored during text typing process.


Figure 1: Typical Chinese Pinyin input method (Sogou-Pinyin).

The typical way to type in Chinese words is in a sequential manner (Wang et al., 2001). iRearch (2009) showed that Pinyin input methods have the biggest share of Chinese speakers. We take one of them for example. Suppose users want to type in Chinese word "今天(today)". Firstly, they mentally generate and physically type in corresponding Pinyin "jintian". Then, a Chinese Pinyin input method displays a list of Chinese homophones, as shown in Figure 1. Finally, users visually search the target word from candidates and select numeric key, e.g. '1'-'9'(<NUM#1>-<NUM#9>) or space key (<SPACE>, a shortcut



*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 662–667, Beijing, China, July 26-31, 2015. 

for numeric key '1') to get the target word (Zheng et al., 2011). Other Chinese input methods, like Wubi, also take these three steps. Typing English words does not involve the last two steps, which indicates that it is on one side more complicated for Chinese users to type in Chinese words than English, but on the other side more convenient for us to obtain additional information produced by users in typing process. We define numeric keys and the space key as **selection keys** for choosing the target word. For sentence "今天天气不错。(Nice weather today.)", one general sequence with selection keys is like "今天(today)<SPACE>天气(weather)<NUM#2>不错(not bad)<SPACE>。" or "今天 (today) <SPACE>天气不错 (weather is not bad) <SPACE>。" In a certain sense, these user-produced selection keys play a role of word delimiters in a very natural way.

In this paper, we propose the concept of Natural Typing Annotations (NTAs) that are potential word delimiters produced by users while typing Chinese words, and verify that it is plausible to automatically generate labeled data for CWS by exploiting NTAs. According to the principle of statistical sampling, texts with NTAs are gathered from 384 users. Specifically, since the ultimate goal is to exploit NTAs to automatically generate labeled data for word segmentation, the main task is to select high-quality NTAs, which largely overlap with gold segmentation. We do this by 1) training a classifier to distinguish acceptable-quality NTAs from low-quality ones, and then 2) using a voting mechanism to further locate the high-quality NTAs among those identified by the classifier in the first step. Experiments show that Support Vector Machine (SVM) and voting mechanism are effective for this work and the high-quality NTAs texts can be used as the training data for improving the performance of supervised word segmentation model in out-of-domain. In addition, some evidence is provided that user-produced NTAs might be particularly useful to deal with out-of-vocabulary (OOV) words.

In the rest of the paper, we briefly introduce the gold standard and baseline segmenter of our work in section 2, then describe the definition and characteristic of natural typing annotations (NTAs) in section 3, and finally elaborate on the strategy of locating high-quality NTAs texts in section 4.After giving the experimental results and analysis in section 5, we come to the conclusion and the implication of future work.

## 2 Gold Standard and Baseline segmenter

There are many different standards for word segmentation, and different tasks usually need different standards. The Sighan Bakeoff uses four well-known standards made by four different organizations: Academia Sinica (AS), City University of Hong Kong (CU), Peking University (PKU), and Microsoft Research (MSR). In this study, we take MSR segmentation standard as **gold standard**. Following the work of Zhao et al. (2010) and Sun and Xu (2011), a Conditional Random Fields (CRF) model (Lafferty et al., 2001) is trained with the training corpus of MSR from Sighan Bakeoff-2, to be a baseline segmenter. This general-purpose segmenter is called as **CRF+MSR** in this paper.

## 3 Natural Typing Annotations Texts

### 3.1 Formulation

A Chinese sentence is represented as $S = c_1 c_2 ... c_N$ ($c_i$ stands for a Chinese character, $N$ is the length of sentence $S$). One of the possible sequences with selection keys is defined as $\pi(S) = | c_1 ... c_{i_1 - 1} | c_{i_1} ... c_{i_2 - 1} | ... | c_{n_1} ... c_N |$.Here, we use the symbol "|" instead of each selection key. "|" is the "**Natural Typing Annotation (NTA)**", which is naturally annotated by users when typing Chinese words. Between the two neighboring "|"s is a **segment**. Then the user-produced $\pi(S) = | segment_1 | segment_2 | ... | segment_M |$ ($M \le N$, $M$ is the number of segments in sentence $S$) is called as **NTAs text** or **NTAs corpus**.

### 3.2 Collection of NTAs Texts

We need to collect user-produced NTAs texts independently because there are no similar or alternative open corpora. We posted a public notice on the Internet to gather volunteer participants. For comparison, they were told to type in the same assigned test text while our software recorded the character sequence with NTAs. Two explanations are given as followed. First, to get more users' feedback and keep the significance level of the experiment, we only have 365 Chinese characters in the test text, which contains words with ambiguous meaning, named entities (NEs), neo-logisms and typo-prone words. Even the state-of-the-art segmenters cannot handle this test text very well. Second, according to statistical sampling theory, if we want

a 95% confidence interval to have a margin of error less than 5%, the sample size should be no less than 384. Therefore, we randomly accept 384 volunteers to join our typing experiment and get user-produced NTAs texts from them.

### 3.3 Analysis of Collected NTAs Texts

Users' overall typing habit can be drawn through the analysis of the collected NTAs texts. We firstly focus on segment, because it is the basic unit in our texts. A total of 66,232 segments are obtained from all texts, but only 883 of them are not repeated. Using $Length(seg)$ to represent the length of a segment is easy to get a frequency distribution of different $Length(seg)$ and find that the length of frequent segments is largely concentrated during 1 to 4. The same statistics can be conducted separately with the word segmentation results by gold standard and CRF+MSR. We use relative frequencies to illustrate the overall trend of three results, as shown in Figure 2.

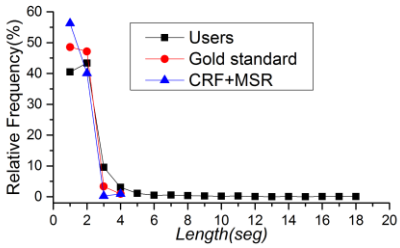


Figure 2: Relative frequencies of segment length from three segmentation.

The results suggest that most Chinese speakers are reluctant to put a long text string into one segment, which is roughly consistent with behavioral economics and psycho-linguistic. Users consciously avoid the mistakes that might be brought by typing in long sequence at a time. Besides, people seldom put illogical sequence of characters into 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. Consequently, the constitution of segment is a reflection of natural language logic.

## 4 High-quality NTAs Texts

### 4.1 User's Typing Patterns

In this section, we investigate the collected NTAs texts at the sentence level. Direct visual impression is that 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 assessed with 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 segmentation of $S_1$, and others are representative sequences from 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|"

We discover three typing patterns of users. The first one is **Discrete Pattern**, where the characters belonging to one segment in the light of gold standard are separated into several segments, such as $\pi_5(S_1)$. The second is **Adhesive Pattern**, which suggests that two or more adjacent individual words by gold standard come together to form one segment, like $\pi_3(S_1)$ and $\pi_4(S_1)$. The third is **Acceptable Pattern**, where user-produced segmentation is largely or exactly the same with the gold standard, such as $\pi_1(S_1)$ and $\pi_2(S_1)$. We find that discrete pattern and adhesive pattern are useless for word segmentation. So we call those NTAs texts that follow acceptable pattern **acceptable-quality NTAs texts**, and others low-quality ones. Furthermore, among acceptable-quality NTAs texts, some of them are more close to gold standard, which is called as **high-quality NTAs texts**. Our strategy is 1) to use a classifier to find all acceptable-quality NTAs texts, and then 2) to further locate the high-quality NTAs texts among those identified by the classifier in the previous step.

### 4.2 The Classification Approach

Identification of acceptable-quality NTAs texts is a typical binary classification problem. Effective and logical features should be identified to model a classifier. We select the following five features because they are simple but outstanding against other alternatives for this work.

$$Features = \begin{Bmatrix} \text{Len}, \\ \text{SegNum}, \\ \text{SingleSegNum}, \\ \text{MaxConSingleSegNum}, \\ \text{MaxSegLen} \end{Bmatrix}$$

**Len** is the abbreviation for length of a sentence, and **SegNum(SN)** stands for the number of the segments in a sentence. These two features can be used to determine whether the percentage of character number of a sentence and the segment number of a sentence is in a proper range.

**SingleSegNum(SSN)** stands for the number of the segments whose length equals 1 in a sentence. **MaxConSingleSegNum(MCSSN)** is the maximum number of continuous segments whose length is 1. **MaxSegLen(MSL)** means the length of segment with most characters. These three features can be used to identify whether discrete or adhesive phenomena prevail in a sentence.

### 4.3 The Voting Mechanism

As the classification approach brings lots of acceptable-quality NTAs texts, voting mechanism is introduced to further locate the high-quality NTAs texts. For a sentence $S_i$, there possibly exist different user-produced segmentations $\pi_1(S_i)$, $\pi_2(S_i)$, … , $\pi_k(S_i)$ ($k$ is the total number of these segmentations). If $\pi_j(S_i)$ appears in different users' texts, these texts practically vote for $\pi_j(S_i)$. Different users' texts practically vote for $\pi_j(S_i)$, which appears in these texts. Thus every sentence $S_i$ in a text can get a score:

$$SCORE_{\pi_j(S_i)} = \log_2 count(\pi_j(S_i)) \quad (1)$$

$count(\pi_j(S_i))$ calculates how many users input $S_i$ with segmentation $\pi_j(S_i)$. A text (namely a user) also has a score:

$$SCORE_{text} = \frac{\sum_{\pi_j(S_i) \in text} \log_2 count(\pi_j(S_i))}{num_{\pi_j(S_i) \in text}} \quad (2)$$

$num_{\pi_j(S_i) \in text}$ is the number of sentences in this text.

This score helps us to identify high-quality NTAs texts from all acceptable-quality ones.

## 5 Experiments

### 5.1 Identification of High-quality NTAs Texts

In this experiment, we verify the effectiveness of classifier and voting mechanism on locating high-quality NTAs texts from 384 collected ones. You can download part of our collected texts from *https://github.com/dakuiz/NTAs*.

#### 5.1.1 The Classification Experiment

We randomly select 32 NTAs texts that contain 1,089 sentences, and then manually label them to form training set. Taking $S_1$ mentioned in 4.1 as an example, the manual-labeled training data are shown in table 1. The label 1 and 0 represent acceptable-quality and low-quality NTAs sentence separately.

| | Len | SN | SSN | MCSSN | MSL | label |
|---|---|---|---|---|---|---|
| $\pi_1(S_1)$ | 16 | 8 | 2 | 1 | 3 | 1 |
| $\pi_2(S_1)$ | 16 | 11 | 6 | 3 | 2 | 1 |
| $\pi_3(S_1)$ | 16 | 5 | 2 | 1 | 5 | 0 |
| $\pi_4(S_1)$ | 16 | 2 | 0 | 0 | 11 | 0 |
| $\pi_5(S_1)$ | 16 | 15 | 14 | 12 | 2 | 0 |

Table 1: examples of training data for classifier.

Package of libSVM (Chang and Lin, 2011) is used here. Radial basis function is adopted as the kernel function where gamma value is set 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. Ten runs are performed with each run using a different part as the testing set. It is conducted ten times and every part should be testing set once. Classification accuracy of the experiment is listed in the table 2.

| Num | Accuracy(%) |
|---|---|
| 1 | 96.33 |
| 2 | 97.22 |
| 3 | 97.25 |
| 4 | 97.25 |
| 5 | 89.91 |
| 6 | 98.17 |
| 7 | 94.50 |
| 8 | 94.59 |
| 9 | 94.55 |
| 10 | 98.11 |
| **Average** | 95.79 |

Table 2: 10-fold cross validation results.

Since the results indicate the validity of our classification approach, we use this classifier to handle collected NTAs texts. If 85% of sentences in a text are acceptable-quality, we select this text as acceptable-quality NTAs text. Finally, we obtain 211 acceptable-quality NTAs texts from all 384 collected ones.

#### 5.1.2 The Voting Experiment

According to voting mechanism in section 4.3, every acceptable-quality NTAs text can get a score to rank itself. Table3 shows top three high-quality NTAs texts with their user-produced word segmentation results compared with that of CRF+MSR. Because CRF+MSR is a general-

purpose segmenter and test data does not come from news wire, its performance drops significantly in out-of-domain.

Table 3 suggests that high-quality NTAs texts are very close to gold standard of word segmentation. To discover the causes of errors, we manually inspected these three texts and found the major error is adhesive phenomenon between simple words. For example, gold segmentation "|这|几|款|" is formed as "|这几款|" by users. This is an error in word segmentation competition, but in some application scenarios, like machine translation, "|这几款|"is better than "|这|几|款|". Similar phenomena shed light on understanding what a "word" really is.

| **Word segmentation from** | ***p*** | ***r*** | ***f*** | $r_{oov}$ |
|---|---|---|---|---|
| CRF+MSR | 90.86 | 92.02 | 91.43 | 50.00 |
| Text#top1 | **92.82** | 90.19 | **91.49** | **100.00** |
| Text#top2 | **91.50** | 88.29 | 89.87 | **100.00** |
| Text#top3 | 90.38 | 87.33 | 88.83 | **100.00** |

Table 3: Test text word segmentation results from general-purpose segmenter and top 3 texts.

### 5.2 Effectiveness of High-quality NTAs Corpus on Improving Word Segmentation

It is generally agreed among researchers that users' behavioral patterns maintain consistent over a long period of time (Zhang et al., 2013; Stephane, 2009). In table 3, we listed top 3 high-quality NTAs texts. Users who generated these three NTAs texts are stable sources to provide more well-segmented texts.

To evaluate the effectiveness of high-quality NTAs corpus on building training data for segmenter, we use a web crawler to get 40k Micro-blog (weibo.com) corpus and randomly divided it into 4 equal shares, i.e. A, B, C, T text. The provider of top1 text is invited to retype A text to produce A NTAs text. B and C NTAs texts are separately obtained from other two providers. We use A, B, C NTAs texts as training data to get a CRF segmentation model, which is called as **CRF+NTAs**. Then we train anther CRF segmenter with a combination of A, B, C NTAs texts and the training corpus of MSR from Bakeoff-2, called as **CRF+MSR+NTAs**. We select 1,000 sentences from T text to manually segment by gold standard, and use them to form our test set that contains 6528 characters. The results of the three segmenters on this Micro-blog test set is shown in table 4.

The model directly trained by Micro-blog high-quality NTAs corpus is better than general-purpose segmenter but far from the model trained by the combination of MSR and Micro-blog high-quality NTAs corpus. This is the most compelling evidence to show that high-quality NTAs corpus can be used for improving word segmentation model in out-of-domain.

| **Word segmentation from** | ***p*** | ***r*** | ***f*** |
|---|---|---|---|
| CRF+MSR | 88.95 | 90.63 | 89.78 |
| CRF+NTAs | **92.38** | 89.76 | **91.05** |
| CRF+MSR+NTAs | **96.27** | **94.83** | **95.54** |

Table 4: Segmenters' results on test data.

We also find out that the NTAs might be particularly useful to identify OOV words, such as proper names and neo-logisms. If users frequently put some characters in one segment, this segment may be some new word or the new internet slang, such as "白富美(white, rich and pretty) ", "萌萌哒(very cute)", "十动然拒(someone is moved but refuses to become girl/boyfriend)", etc.

## 6 Conclusion and Future Work

In this paper, we investigate Natural Typing Annotations (NTAs) that are potential word delimiters generated by Chinese speakers while typing Chinese words. The effectiveness of high-quality NTAs corpus on improving word segmentation is evaluated.

Though it is convenient for users to read, sequence of pure characters, namely without any recorded delimiters produced by inputters, loses lots of valuable information, e.g. NTAs. We strongly recommend that NTAs can be recorded in an invisible manner for normal users by dominant text editors, such as MS Word, Notepad, vi, emacs, etc.

In future, we will: 1) collect more NTAs texts from various users; 2) do further work on how to fully leverage NTAs to improve word segmentation; 3) call for dominant text editors to record NTAs.

## Acknowledgments

Sincere thanks to the three anonymous reviewers for their thorough reviewing and valuable suggestions! The authors were supported by the National Strategic Basic Research Program ("973" Program) of the Ministry of Science and Technology of China (No. 2012CB720702 and 2013CB329303) and National Science Foundation of China (No. 61303105).

## Reference


Chih-Chung Chang, Chih-Jen Lin. 2011. *LIBSVM: A library for support vector machines*. In TIST.

Yuan Chen. 1997.*Chinese Language Processing*. Shanghai Education publishing company.

iRearch.2009. *2009 China Desktop Software Development Research Report*. http://report.iresearch.cn/1290.html.

Wenbin Jiang, Liang Huang, Qun Liu. 2009. *Automatic adaptation of annotation standards: Chinese word segmentation and pos tagging – a case study*. In ACL-AFNLP.

John D. Lafferty, Andrew McCallum, Fernando C. N. Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In ICML.

Zhongguo Li, Maosong Sun. 2009. *Punctuation as Implicit Annotations for Chinese Word Segmentation*. Computational Linguistics.

Wenbin Jiang, Meng Sun, Yajuan Lü, Yating Yang, Qun Liu. 2013. *Discriminative Learning with Natural Annotations: Word Segmentation as a Case Study*. In ACL.

Yang Liu, Yue Zhang. 2012. *Unsupervised domain adaptation for joint segmentation and POS-tagging*. In COLING.

Yijia Liu, Yue Zhang, Wanxiang Che, Ting Liu, Fan Wu. 2014. *Domain Adaptation for CRF-based Chinese Word Segmentation using Free Annotations*. In EMNLP.

Xipeng Qiu, ChaoChao Huang, Xuanjing Huang. 2014. *Automatic Corpus Expansion for Chinese Word Segmentation by Exploiting the Redundancy of Web Information*. In COLING.

Weiwei Sun, Jia Xu. 2011. *Enhancing chinese word segmentation using unlabeled data*. In EMNLP.

Lucas Stephane. 2009. *User Behavior Patterns: Gathering, Analysis, Simulation and Prediction*. In HCI.

Jingtao Wang, Shumin Zhai, Hui Su. 2001. *Chinese input with keyboard and eye-tracking: an anatomical study*.In CHI.

Yiou Wang, Jun'ichi Kazama, Yoshimasa Tsuruoka,Wenliang Chen, Yujie Zhang, Kentaro Torisawa. 2011. *Improving chinese word segmentation and pos tagging with semi-supervised methods using large auto-analyzed data*. In IJCNLP.

Nianwen Xue, Libin Shen. 2003. *Chinese word segmentation as lmr tagging*. In SIGHAN.

Chunhong Zhang, Yaxi He, Yang Ji. 2013. *Temporal Pattern of User Behavior in Micro-blog*. In JSW.

Yue Zhang, Stephen Clark. 2007. *Chinese segmentation with a word-based perceptron algorithm*. In ACL.

Hai Zhao, Chang-Ning Huang, Mu Li, Bao-Liang Lu. 2010. *A unified character-based tagging framework for chinese word segmentation*. In ACM.

Yabin Zheng, Lixing Xie, Zhiyuan Liu, Maosong Sun, Yang zhang, Liyun Ru. 2011. *Why Press Backspace? Understanding User Input Behaviors in Chinese Pinyin Input Method*. In ACL.