# An Agent-Based Approach to Chinese Word Segmentation

**Samuel W.K. Chan**
Dept. of Decision Sciences
The Chinese University of Hong Kong
Hong Kong, China
swkchan@cuhk.edu.hk

**Mickey W.C. Chong**
Dept. of Decision Sciences
The Chinese University of Hong Kong
Hong Kong, China
mickey@baf.msmail.cuhk.edu.hk

## Abstract

This paper presents the results of our system that has participated in the word segmentation task in the Fourth SIGHAN Bakeoff. Our system consists of several basic components which include the pre-processing, token identification and the post-processing. An agent-based approach is introduced to identify the weak segmentation points. Our system has participated in two open and five closed tracks in five major corpora. Our results have attained top five in most of the tracks in the bakeoff. In particular, it is ranked first in the open track of the corpus from Academia Sinica, second in the closed track of the corpus from City University of Hong Kong, third in two closed tracks of the corpora from State Language Commission of P.R.C. and Academia Sinica.

## 1 Introduction

Our word segmentation system consists of three major components, namely, the pre-processing, token identification and the post-processing. In this paper, an overview of our system is briefly introduced and the structure of the paper is as follows. Section 2 presents the system description. An agent based approach is introduced in the system. Associated to each agent in the system, a vote is cast to indicate the certainty by each agent in the system. In Section 3, we describe the experimental results of our system, followed by the conclusion.

## 2 System Description

### 2.1 Preprocessing

In the preprocessing, the traditional Chinese characters, punctuation marks and other symbols are first identified. Instead of training all these symbols with the traditional Chinese characters in an agent-based system, an initial, but rough, segmentation points ($SP_r$) are first inserted to distinguish the symbols and Chinese characters. For example, for the input sentence shown in Figure 1, segmentation points are first assumed in the sentence as shown in the Figure 2, where '/' indicates the presence of a segmentation point. This roughly segmented sentence is then subject to an agent-based learning algorithm to have the token identification.

昨日6時05分終於成功發射了第一顆自行研製的探月衛星「嫦娥一號」。

Figure 1: Original sentence for the process

昨日/ 6/ 時/ 05/ 分終於成功發射了第一顆自行研製的探月衛星/「/ 嫦娥一號/ 」/ 。

Figure 2: Rough segmented sentence from preprocessing

### 2.2 Token Identification

In this stage, a learning algorithm is first devised and implemented. The algorithm is based on an agent based model which is a computational model for simulating the actions and interactions of an orchestra of autonomous agents in the determination of the possible segmentation points ($SP_l$) (Weiss, 1999; Wooldridge, 2002). Each agent will make its own decision, i.e., either true or false, for the insertion of "/" between the two characters. Moreover, associated with each decision, there is a vote that reflects the certainty of the decision. For each training corpus, we have trained more than 200 intelligent agents, each of which exhibits certain aspects of segmentation experience and language behaviors. In making the final verdict, the system will consult all the related agents by summing up their votes. For example, as shown in Table 1, the vote that supports there is a segmentation point between the characters 昨 and 日 is zero while 57.33 votes recommend that there should have no break point. All these votes are logged for the further post-processing.

| $C_1$ | $C_2$ | *Vote(T)* | *Vote(F)* | *ND* | *Outcome* |
|---|---|---|---|---|---|
| 昨 | 日 | 0 | 57.33 | 1.000 | false |
| 分 | 終 | 44.52 | 6.54 | 0.744 | true |
| 終 | 於 | 0 | 57.74 | 1.000 | false |
| 於 | 成 | 64.61 | 0 | 1.000 | true |
| 成 | 功 | 0 | 60.23 | 1.000 | false |
| 功 | 發 | 56.29 | 0.99 | 0.965 | true |
| 發 | 射 | 0.58 | 58.22 | 0.980 | false |
| 射 | 了 | 58.21 | 0 | 1.000 | true |
| 了 | 第 | 57.80 | 0 | 1.000 | true |
| 第 | 一 | 0 | 51.34 | 1.000 | false |
| 一 | 顆 | 48.70 | 0 | 1.000 | true |
| 顆 | 自 | 60.04 | 0 | 1.000 | true |
| 自 | 行 | 0 | 53.97 | 1.000 | false |
| 行 | 研 | 46.19 | 2.00 | 0.917 | true |
| 研 | 製 | 0 | 58.32 | 1.000 | false |
| 製 | 的 | 62.44 | 0 | 1.000 | true |
| 的 | 探 | 59.16 | 0 | 1.000 | true |
| 探 | 月 | 4.89 | 40.81 | 0.786 | false |
| 月 | 衛 | 45.83 | 3.41 | 0.862 | true |
| 衛 | 星 | 0 | 60.91 | 1.000 | false |
| 嫦 | 娥 | 0 | 59.39 | 1.000 | false |
| 娥 | 一 | 54.44 | 0.48 | 0.983 | true |
| 一 | 號 | 11.98 | 27.94 | 0.400 | false |

Table 1: Votes from agents and the ND of the corresponding segment point.

昨日/ 6 / 時/ 05 / 分/ 終於/ 成功/ 發射/ 了 / 第一/ 顆/ 自行/ 研製/ 的/ 探月/ 衛星/「/ 嫦娥/ 一號/ 」/ 。

Figure 3: Segmented sentence based on the votes from all agents.

### 2.3 Post-processing

In our experience, our system is most likely to generate over-segmented sentences. Several techniques have implemented in our post-processing to merge several tokens into ones. As shown in the previous steps, we have introduced two main types of segmentation points, $SP_r$ and $SP_l$. In the type $SP_r$, segmentation points are pre-inserted between symbol and Chinese characters. For example, the token 6 時 will become 6/ 時 in the early beginning. Obviously, this kind of errors should be identified and the segmentation points should be removed. Similarly, in $SP_l$, segmentation points are decided by the votes. Our post-processing is to identify the weak segmentation points which are having tie-break votes. A normalized difference (*ND*) is defined for the certainty of the segmentation.

$$ND = \frac{\left|Vote_{true} - Vote_{false}\right|}{Vote_{true} + Vote_{false}} \qquad \text{Eqn.(1)}$$

The smaller the value of the *ND*, the lesser the certainty of the segmentation point. We define the segmentation point as weak if the value of *ND* is smaller than a threshold. For a weak segmentation point, the system will consult a dictionary and search for the presence of the token in the dictionary. The segmentation point will be removed if found. Otherwise, the system will leave as it is. As shown in the Table 1, almost all segmentation points with the ND value equal to 1. This shows that all the votes from the agents support the same decision. However, it seems that not all agents have the same decision to the last characters pair "一號", with *ND* equal to 0.4. If the threshold is set to be 0.4, the segmentation point will be reexamined in our post-processing.

Our dictionary is constructed by tokens from the training corpus and the local context of the text that is being segmented. That is to say, besides the corpus, the tokens from the previous segmented text will also contribute to the construction of the dictionary. On the other hand, Chinese idiom should be in one token as found in most dictionaries. However, idiom sometimes would be identified as a short phrase and segmented into several pieces. In this case, we tend to merge these small fragments into one long token. On the other hand, different training sources may produce different segmentation rules and, thus, produce different segmentation results. In the open tracks, some handlers are tailor-made for different testing data. These include handlers for English characters, date, time, organization.

昨日/ 6 時/ 05 分/ 終於/ 成功/ 發射/ 了/ 第一/ 顆/ 自行/ 研製/ 的/ 探月/ 衛星/「/ 嫦娥/ 一號/ 」/ 。

Figure 4: Final result of the segmentation

## 3 Experiments and Results

We have participated in five closed tracks and two open tracks in the bakeoff. While we have built a dictionary from each training data set for the closed tracks, a dictionary of more than 150,000 entries is maintained for the open tracks. Table 2 shows the size of the training data sets.

| Source of training data | Size |
|---|---|
| Academia Sinica (CKIP) | 721,551 |
| City University of Hong Kong (CityU) | 1,092,687 |
| University of Colorado (CTB) | 642,246 |
| State Language Commission of P.R.C. (NCC) | 917,255 |
| Shanxi University (SXU) | 528,238 |

Table 2: Size of the training data in the bakeoff.

Tables 3 and 4 show the recall (*R*), precision (*P*), *F*-score (*F*) and our ranking in the bakeoff. All the rankings are produced based on the best run of the participating teams in the tracks.

| | *R* | *P* | *F* | *Rank* |
|---|---|---|---|---|
| CityU | 0.9513 | 0.9430 | 0.9471 | *2nd* |
| CKIP | 0.9455 | 0.9371 | 0.9413 | *3rd* |
| NCC | 0.9365 | 0.9365 | 0.9365 | *3rd* |
| SXU | 0.9558 | 0.9552 | 0.9555 | *5th* |

Table 3: Performance of our system in the closed tracks of word segmentation task in the bakeoff.

| | *R* | *P* | *F* | *Rank* |
|---|---|---|---|---|
| CKIP | 0.9586 | 0.9541 | 0.9563 | *1st* |
| NCC | 0.9440 | 0.9517 | 0.9478 | *4th* |

Table 4: Performance of our system in the open tracks of word segmentation task in the bakeoff.

From the above tables, we have the following observations:

- First, our system is performing well if it is a sufficient large set of training data. This is evidenced by the results found in the training data from CKIP, CityU and NCC.
- Second, the dictionaries play an important role in our open tracks. While we have maintained a dictionary with 150,000 traditional Chinese words, no such a device is for our simplified characters corpora. Certainly, there is a room for our further improvement.

## 4 Conclusion

In this paper, we have presented the general overview of our segmentation system. Even though, it is our first time to participate the bakeoff, the approach is promising. Further exploration is needed to enhance the system.

## Acknowledgement

The work described in this paper was partially supported by the grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. CUHK4438/04H and CUHK4706/05H).

## References

Weiss, G. (1999) *Multiagent Systems, A Modern Approach to Distributed Artificial Intelligence*, MIT Press.

Wooldridge, M. (2002). *An Introduction to MultiAgent Systems*, John Wiley.