Quality Assurance for Segmentation and Tagging of Chinese Novels in the Ming and Qing Dynasties

Dan Xiong[1], Qin Lu[1], Fengju Lo[2], Dingxu Shi[3], Tin-shing Chiu[1]

[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
{csdxiong, csluqin, cstschiu}@comp.polyu.edu.hk
[2]Department of Chinese Linguistics & Literature, Yuan Ze University, Taiwan
gefjulo@saturn.yzu.edu.tw
[3]Department of Chinese & Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong
ctdshi@polyu.edu.hk

***Abstract*—This paper presents a word segmentation and named entity tagging project which annotates Chinese novels in the Ming and Qing dynasties. Computer-aided tools are used to assist the annotation. The focus of this paper will be on the quality assurance measures to ensure precision and consistency. The specification for word segmentation and named entity tagging is formulated based on the standards for modern Chinese segmentation commonly used in Mainland China and in Taiwan as well as the analysis of differences between Chinese classics and modern Chinese. The specification is established through iterative refinements. This refinement process can offer valuable insights into the quality control of computer-aided processing performed on Chinese literature works in the Ming and Qing dynasties and can be applied to those in even earlier periods. The finalized corpus, built in a computer-aided, manually-reviewed method in accordance with the specification, can be used for researches in literature, linguistics, information technology, and teaching of Chinese.**

Keywords—Quality assurance, word segmentation, tagging, named entities, novels in the Ming and Qing dynasties

I. INTRODUCTION

The building of natural language resources is an essential step in natural language processing applications. The quality assurance in building these resources ensures that the language resources can be used with confidence. This paper presents a project to build a high-quality tagged corpus of four classical Chinese novels in the Ming and Qing dynasties, namely, *A Dream of Red Mansions* (《紅樓夢》, Hóng Lóu Mèng), *Romance of the Three Kingdoms* (《三國演義》, Sān Guó Yǎn Yì), *Water Margin* (《水滸傳》, Shuǐ Hǔ Zhuàn), and *The Golden Lotus* (《金瓶梅》, Jīn Píng Méi). The main objective of this project is to help build a large collection of tagged resources for computer-aided lexical semantic analysis of classical Chinese literature. Hopefully the result of this work can provide insights into the transition of Chinese literature from its traditional forms such as traditional verses and vernacular styles to modern Chinese. Based on the objective of the project, the tasks must include word segmentation as well as named entity tagging because named entities play an important part in semantic understanding of Chinese literature. The specification for segmentation and named entity tagging is thus developed for this project.

The rest of the paper is organized as follows. Section II explains the main differences between modern Chinese and Chinese classics, especially in named entities. Section III briefly describes the specification of segmentation and named entity tagging. Section IV presents our quality control process as well as the measures we have taken to improve accuracy and also shows the performance of the tagging work.

II. FEATURES OF NOVELS IN THE MING AND QING DYNASTIES

The Ming and Qing novels are quite different from modern Chinese and thus their word segmentation and named entity tagging require a separate specification. The first task of this project is to analyze the differences between the Ming and Qing novels and modern Chinese to establish the segmentation and tagging specification applicable to this kind of classical literature.

A. Differences from Modern Chinese in Linguistic Aspects

The novels in the Ming and Qing dynasties, written in the vernacular style and developed during the transition from ancient Chinese to modern Chinese, have unique linguistic features. The Ming and Qing novels are distinguished from modern Chinese in style, semantics, and syntax.

First, the Ming and Qing novels, in a more compact writing style, often use short words composed of only a single character. This stands directly in contrast with modern Chinese where two-character to four-character words are dominant. For example, in the four classical novels mentioned in Section I with 2,623,540 characters in total, single-character words are roughly 13 times as many as two-character words, 32 times as many as three-character words, and 65 times as many as other multi-character words. This means that words are used at a much smaller granularity.

Another feature is the chapter-by-chapter style, as they are famously called Zhang-Hui (章回) novels, in which each chapter is headed by a couplet indicating the general meaning of its content and also likely ends with a couplet or poem. The Ming and Qing novels contain a large number of poems, couplets, prose poems, lyrics, etc., which are written in classical style. They are not considered for segmentation in this work because processing them requires quite a different way of treatment.

The Ming and Qing novels also differ from modern Chinese in the use of lexical entries. In the Ming and Qing novels, many lexical entries are partially or completely different from those in modern Chinese. For instance[1], in the Ming and Qing novels, "色色" (sè sè) means "every kind of", but in modern Chinese the equivalent is "樣樣" (yàng yàng). It is worth mentioning that "足下" (zú xià) may be used in both the Ming and Qing novels and modern Chinese. Yet, their meanings are quite different. In the Ming and Qing novels, it is a respectful form of addressing a person of equal rank; in modern Chinese, it is not used to address a person. Some lexical entries in the Ming and Qing novels are no longer used in modern Chinese, especially those describing clothes, hair ornaments, decorations, household utensils, etc.

B. Differences from Modern Chinese in Named Entities

Named entities in the Ming and Qing novels are quite different from those in modern Chinese. In this work, named entities most commonly used in the Ming and Qing novels are classified into six categories: personal name (人名 rén míng), term of address (人物稱謂 rén wù chēng wèi), name of official position (官職 guān zhí）and title of nobility or honour (爵位 jué wèi，封號 fēng hào), place name (地名 dì míng), building name (建築名 jiàn zhù míng), and organizational name (組織名 zǔ zhī míng).

In the old time, a person, especially if he has certain social status, may have many names besides his surname and given name. An adult male usually has a courtesy name (字 zì) as a symbol of adulthood and respect. A person may also have one or more art names (號 hào), which are alternative courtesy names mostly with three or four characters. After death, an honorable person may be awarded a posthumous name (謚號, shìhào), and an emperor is given a temple name (廟號, miào hào). Term of address, an important part of everyday communication, is widely used in the Ming and Qing novels. Under different settings, a person can be addressed by his given name, courtesy name, nickname, etc. or combination of any form of the above names with a title indicating the person's rank or official position. Therefore, tagging terms of address becomes a complicated but meaningful job.

Reference to geographic locations and administrative divisions changes over time for one reason or the other, which can introduce a lot of location names that are different from those in modern China. Their annotation is important if linkage of the past and the present needs to be identified in the research of literature and history.

III. SPECIFICATION FOR WORD SEGMENTATION AND NAMED ENTITY TAGGING

Based on the standards for modern Chinese segmentation widely accepted in Mainland China [1] and in Taiwan [2] as well as the analysis on the differences between the Ming and Qing novels and modern Chinese, this work has formulated the basic segmentation and tagging rules applicable to the Ming and Qing novels. Then some sample texts were segmented and tagged according to these rules to verify their feasibility and flexibility. The specification has been established iteratively through review of the tagging work.

First of all, a segmentation unit is considered as the smallest linguistic unit that "has specific semantic or grammatical functions" according to GB13715, i.e., China's national segmentation standard for modern Chinese information processing [3]. From the perspective of semantics, the basic principle is to segment text strings into the smallest units without semantic loss, distortion, or ambiguity. We also set out some rules from the perspective of syntax as complementary guidelines.

As mentioned earlier, a variety of compound named entities can be found in the Ming and Qing novels. For example, a term of address may be composed of any form of the name and title, while any name, term of address, or title may also be used as a part of a place name or a building name. The project has established a set of refined tagging rules for different kinds of named entities to ensure that they can be segmented and labeled in a consistent and flexible way. In general, the project follows the approach of Peking University [1] for named entity tagging: square brackets ([]) are used to enclose compound named entities and labels are marked by the slash sign (/). The units in the square brackets are segmented and tagged according to the unified specification. For example, the compound building name "[[忠靖/侯]/nu2 史/nr1 府]/nv1" ([[Zhōngjìng/Hóu]/nu2 Shǐ/nr1 fǔ]/nv1), which is composed of the title of nobility "[忠靖/侯]/nu2" ([Zhōngjìng/Hóu]/nu2, Marquis of Zhōngjìng) and the surname "史/nr1" (Shǐ/nr1), refers to the residence of Marquis of Zhōngjìng whose surname is Shǐ.

For literature/historical study, it is important to distinguish real persons and places from fictional ones. Usually, persons and places in novels are fictional. However, in Chinese classics, there are many references to real persons and places, which can be found in the records of official history books. Such distinction would be helpful in literature/historical research. Our work further classifies personal names, terms of address, place names, and building names into four categories: real entities, mythical entities, fictional entities cited from other literature works, and fictional entities in the novels.

Once the named entities are properly categorized and tagged, not only the correlation between the in-text entities but also the connection between the entities in the Ming and Qing novels and those in other literature works can be established.

The details of the specification were published separately at the 13th Chinese Lexical Semantics Workshop (CLSW2012).

IV. QUALITY CONTROL PROCESS AND PERFORMANCE EVALUATION

A. Overall Process

The four classics are, in principle, tagged manually by a single annotator. An early stage test indicates that pure manual tagging is labor intensive and very error prone. To ensure accuracy, consistency, and compliance with the specification, which are the three major factors to assess the quality, computer-aided tools are used to improve quality. As shown in Figure 1, the project follows a strict quality control process to minimize errors. The process consists of three stages with quality control measures.

In Stage 1, a small sample of text is manually annotated to train the system. The annotation and some relevant classical literature lexicons are then integrated into an existing

[1] The examples given in this paper are all cited from *A Dream of Red Mansions* and *Romance of the Three Kingdoms*.

segmentor/tagger [4], which was originally designed to process modern Chinese. The smallest granularity is used in segmentation. This is to ensure that the system can be adapted to processing classical novels and generate a text with higher quality. Then, in Stage 2, an annotator manually tags named entities and reviews the output of the segmentor/tagger to correct any mistake in strict accordance with the specification. After the text is annotated and reviewed, a tool is used in Stage 3 to check some errors of tags, especially inconsistency. The specification and lexicon entries are updated during Stages 2 and 3 iteratively and then are integrated into the segmentor/tagger and original data.

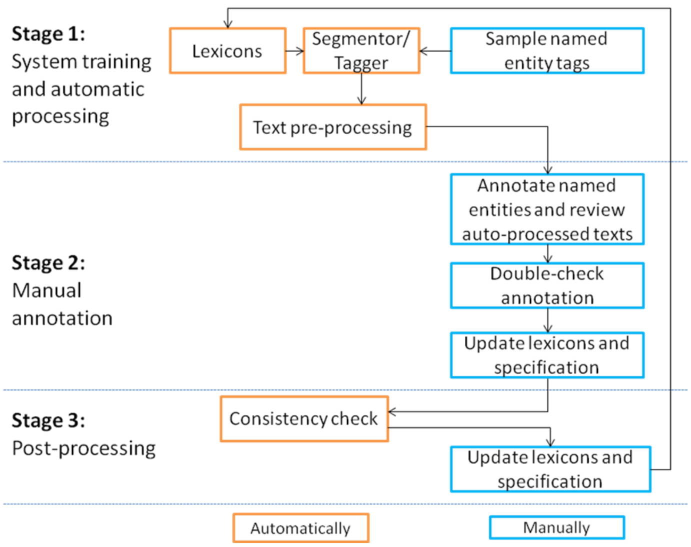

Figure 1 Flow chart of the quality control process

B. *Stage 1: System Training and Automatic Processing*

In this stage, a small set of text is selected for segmentation and named entity tagging. This helps to identify commonly used tags and lexicon entries, and then they are incorporated into the segmentor/tagger originally designed for modern Chinese so that it can be adapted for Chinese classics. With the added knowledge, the segmentor/tagger is used to process the text automatically. For the convenience of lexicon maintenance and minor revision of the specification document, processing of text is done ten chapters at a time.

C. *Stage 2: Manual Annotation*

The task in this stage is to tag all named entities and correct mistakes made during automatic processing. The most challenging work is to treat named entities and lexical entries that can only be interpreted and corrected based on semantics and context, which automated tools cannot handle.

In fact, correct word segmentation can be achieved only through manual work (and in our case with the help of computer-aided tools) because correct output requires the understanding of the semantics which cannot be done in a pipelined approach by the segmentation tool alone. Taking the verb-object structure as an example, verb-object word groups are normally segmented because both the verb and the object usually have independent semantic functions, for instance, "擷/花" (xié/huā, to pick flowers). However, there are many special situations. The following gives some examples to show the different types of cases that require manual work.

- Fixed expressions and the verb-object word groups with a particular meaning that cannot be inferred from the meaning of each separate word are not segmented, for example, "修方" (xiū fāng, to make up prescription).
- Sometimes, the same lexical entry should be treated differently because it may have different grammatical functions. For example, when "差人" (chāi rén) means "to send sb. on an errand", it is segmented into two units. However, when it is used as a noun which usually means "an official servant', it should not be segmented.
- The same lexical entry that has the same grammatical function may have different semantic meanings. For example, "下馬" (xià mǎ) functions as a predicate but carries more than one meaning, including "to dismount from a horse" and "to assume a post", the two most common meanings in the Ming and Qing novels. When it means "to dismount from a horse", it is segmented. When it means "to assume a post", it is not segmented because it has a particular meaning that is different from the ordinary meanings of the individual words.

According to the specification, idioms, set phrases, and other lexical entries with particular meanings different from literal combinations of individual characters are not segmented. Some of these entries are included in the lexicons so that the system will not segment them. But others need to be modified manually and added to the lexicons, which makes the work of this stage necessary. In addition, some set phrases used in Ming Qing novels may not be used in modern Chinese any longer, for example, "每日家" (měi rì jiā, day in and day out).

In terms of named entity tagging, two major problems should be addressed in this stage:

- The categories of named entities are semantic – As discussed in Section III, personal names, terms of address, place names, and building names are further classified into four categories: real entities, mythical entities, fictional entities cited from other literature works, and fictional entities in the novels. This is semantic knowledge and it requires the annotator to have a good understanding of the literature to distinguish real ones from fictional ones with the help of reference books. In this work, *The Twenty-Four Histories* (《二十四史》, Èr Shí Sì Shǐ) are used as references because they are considered as important and credible sources.
- Segmentation and tagging of named entities are context sensitive - As it is commonly known, segmentation of Chinese in general is context dependent. In practice, the tools are trained to handle certain cases in a context sensitive way. However, tagging of some named entities can only be judged manually according to the context, which cannot be identified by software tools. For example, when "兄弟" (xiōng dì) refers to a specific elder or younger brother, it is labeled as the term of address to a person; yet when it serves as a common noun referring to brothers, it should not be labeled in our work.

D. *Stage 3: Post-processing*

The manual annotation and review in Stage 2 can correct most of the problems made by the segmentor/tagger. However, new errors may be left in the modified text, such as missed closing brackets, duplicated opening brackets, missed slash sign, inconsistent tags for the same named entity, etc. In this stage, a consistency check tool is used to identify these kinds of errors. Further errors might be found in the later process of lexicon maintenance. If the problem is due to the lack of certain lexicon entries, they will be added to the lexicons by the tool to avoid future problems. Sometimes, inconsistency may occur because of a previously unseen rule. In this case, the specification document will also be updated to reflect this issue.

E. Performance Evaluation

To ensure quality output, manual annotation is conducted in Stage 2 at least twice by one single annotator who is familiar with Chinese classics. This can reduce the risk of inter-person inconsistency which may arise from manual tagging by different annotators. This issue is more serious for Chinese classics which can only be handled by an annotator with acquired Chinese classics knowledge.

The evaluation of the project consists of two parts. The first part evaluates how effective the segmentor/tagger is and the second part evaluates the quality of the final work. For the first evaluation, ten manually tagged chapters with a total length of 67,181 characters from *A Dream of Red Mansions* that have gone through the three-stage process are used as the golden answer. Then, the output of Stage 1 is evaluated. Results show that the output from using only the automated tools after system training in Stage 1 can achieve 90.14% in precision and 94.48% in recall. This means that without manual check, the quality is not good enough as a resource. For instance, "從/小路/逕/至" (cóng/xiǎo lù/jìng/zhì, to go straight along the footpath)" might be automatically treated as "從小/路/逕/至" by the system because "從小" (cóng xiǎo, from childhood) is a fixed lexical entry in modern Chinese. But in this context, it is treated wrong by the system and should be corrected manually.

In the second part of the evaluation, the manually tagged text that has gone through Stage 3 is checked. Due to limited resources, only ten paragraphs are randomly selected for evaluation, including five paragraphs from the ten chapters used for the first part of evaluation and another five paragraphs randomly selected from the remaining chapters of *A Dream of Red Mansions*. The results are further reviewed independently by a second person, that is, the programmer who has developed the consistency check tool and is quite familiar with the specification. Any inter-person inconsistency will be discussed for confirmation. Results show that the precision of the final output is 99.44%. The errors are further analyzed and categorized as follows:

- About 55.36% of all the errors (0.31% of the whole data) are caused by different interpretations on the semantic meanings of some lexical entries or different understandings of the specification. This reflects the needs to further refine the specification to remove ambiguity in the interpretation of the specification. These errors in principle can be further reduced as the work becomes more mature.
- The remaining 44.64% of errors (0.25% of all data) are human errors which are difficult to be completely removed.

This error analysis indicates that the ceiling of the best possible quality of the data can reach 99.75%. In practice, precision of 99.5% should be an achievable goal which the project is targeting at.

V. Conclusion

This paper explains the quality assurance process for segmentation and named entity tagging of Chinese novels in the Ming and Qing dynasties and presents the measures taken in system training, manual handling, and post-processing. In this paper, the performance of each stage is given to show the effectiveness of the specification and quality assurance measures. The specification is established based on the need of the project and the consideration of the unique features of the Ming and Qing novels compared to modern Chinese, including the chapter-by-chapter style of the novels, the classification of named entities, the use of address terms, and idiomatic expressions. In terms of segmentation, the fundamental principle is to segment text strings into the smallest linguistic units with independent semantic or grammatical functions while meaning loss, distortion, or ambiguity will not be caused. All kinds of named entities, especially the compound ones consisting of several elements, are segmented and tagged in a consistent and flexible way. The tagged corpus, which is expected to have a precision rate over 99.5%, can be applied to different fields such as literature, linguistics, information technology, teaching of Chinese, etc. Co-reference annotation will be conducted in the future.

ACKNOWLEDGEMENT

This paper is partially supported by the Chiang Ching-kuo Foundation for International Scholarly Exchange under the project "Building a Diachronic Language Knowledge-base" (RG013-D-09).

References

[1] S. W. Yu, H. M. Duan, X. F. Zhu, B. Swen, and B. B. Chang, "Specification for Corpus Processing at Peking University: Word Segmentation, POS Tagging and Phonetic Notation (in Chinese)," Journal of Chinese Language and Computing, 13 (2), 2003, pp. 121-158.

[2] Segmentation Principle for Chinese Language Processing (CNS14366) (in Chinese). National Bureau of Standard, Taiwan, 1999.

[3] Y. Liu, Q. Tan, and X. K. Shen, Segmentation Standard for Modern Chinese Information Processing and Automatic Segmentation Methodology (GB13715) (in Chinese). Beijing: Qinghua University Press, 1994.

[4] Q. Lu, S. T. Chan, R. F. Xu, T. S. Chiu, B. L. Li, and S. W. Yu, "A Unicode based Adaptive Segmentor," Journal of Chinese Language and Computing, 14 (3), 2004, pp. 221-234.