Specification for Segmentation and Named Entity Annotation of Chinese Classics in the Ming and Qing Dynasties

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

[1] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
{csdxiong,csluqin,cstschiu}@comp.polyu.edu.hk,
csclaireli@gmail.com
[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. The quality of text segmentation and annotation plays a significant role in Natural Language Processing especially in downstream applications. This paper presents the specification for word segmentation and named entity annotation targeted for novels in the Ming and Qing dynasties. The purpose of this work is to build the foundational work for computer-aided lexical semantic analysis of classical Chinese literature, especially the transition of Chinese literature from its traditional forms such as traditional verses and vernacular styles to modern Chinese. To assist in literature study, an elaborate named entity annotation scheme is specially developed for classical Chinese. Computer-aided segmentation and named entity annotation are conducted on some famous Ming and Qing Chinese classics. The specification for the segmentation and annotation is produced based on the studies of the morphology and semantics differences as well as similarities between classical Chinese and modern Chinese with reference to the existing standards for modern Chinese processing widely used in Mainland China and Taiwan.

Keywords: segmentation and PoS principles, named entities, novels in the Ming and Qing dynasties, computer-aided annotation, semantic analysis.

1 Introduction

With the rapid development of information technology and the digitization of documentation, different kinds of annotated text corpora are established to assist in natural language processing applications. The project "Building a Diachronic Language Knowledge-Base" aims to build a comprehensive knowledge base of Chinese language in different eras ranging from traditional verse (韻文) and vernacular literature (語體文) to modern Chinese. This involves the integration of texts of different styles in different historical periods with annotated information to assist in the understanding

D.H. Ji and G.Z. Xiao (Eds.): CLSW 2012, LNAI 7717, pp. 280–293, 2013.

and computer processing of lexical meaning, semantic marking and classification of semantic concepts, as well as the description of grammatical knowledge. In this way, the knowledge-base is able to make use of computer technology to help semantic, syntactic and discourse analysis as well as subject extraction and classification.

This paper describes the development of the specification for segmentation and named entity annotation for the novels in the Ming and Qing dynasties. The novels of this period, written in vernacular style, are regarded as important evidence in the transition from classical to modern Chinese. The establishment of the specification for word segmentation and named entity annotation is vital for building a high-quality natural language corpus to shed light on the gradual change of Chinese writing and possibly provide links between ancient Chinese and modern Chinese. The specification is designed based on the analysis of existing segmentation and annotation principles established for modern Chinese as well as the differences and similarities between the Ming and Qing novels and modern Chinese. In accordance with the specification, a corpus of four classical novels in the Ming and Qing dynasties, namely, *A Dream of Red Mansions* (《紅樓夢》), *Romance of the Three Kingdoms* (《三國演義》), *Water Margin* (《水滸傳》), and *The Golden Lotus* (《金瓶梅》), will be annotated through a computer-aided method followed by manual review. The result of the work is expected to offer valuable insights into lexical semantic analysis of classical literature in the Ming and Qing dynasties and possibly provide basis for the studies of the evolution of Chinese from ancient time to modern era.

This paper is organized as follows. Section 2 presents the design concepts of the specification and briefly explains the differences between this work and a previous work done by Academia Sinica. Section 3 and 4 describe the principles of word segmentation and named entity annotation targeted for the Ming and Qing novels, respectively. Section 5 discusses the quality assurance process and the test results of annotation evaluation in different stages. Section 6 gives a conclusion.

2 Design Concepts of the Specification

In the development of this specification, the existing standards for modern Chinese segmentation are used as references, including the specification for corpus processing developed by Peking University in Mainland China [1] and the standard of Chinese segmentation used in Taiwan [2]. But, in view of the features of the Ming and Qing novels, it is necessary to establish a segmentation and annotation specification applicable to this kind of classical literature. Compared to modern Chinese, the most distinctive features of the Ming and Qing novels include frequent use of single-character words, the chapter-by-chapter style, lexical meaning, and the use of named entities, which are the key elements to be considered during the design of the specification.

First, the Ming and Qing novels, appearing more compact, use much more single-character words than multi-character words. According to the statistics on the four classical novels mentioned in Section 1, the number of single-character words is 13 times that of two-character words, 32 times that of three-character words, and 65 times that of other multi-character words. The chapter-by-chapter style is another

feature of the Ming and Qing novels, in which each chapter is headed by a couplet implying the main plot of this chapter and also usually ends with a couplet or poem. The novels contain a large number of poems, couplets, prose poems, lyrics, etc., which will be treated later according to the specification for segmenting traditional verse. It is obvious that many lexical entries used in the Ming and Qing novels are partially or completely different from those in modern Chinese. For instance[1], in the Ming and Qing novels, "一心一計"[2] (yī xīn yī jì, wholeheartedly) is used more frequently, but in modern Chinese the equivalent is "一心一意" (yī xīn yī yì).

In addition to the linguistic features, named entities are another distinguishing feature of the Ming and Qing novels. Taking personal names as an example, in ancient China, an adult male may have a courtesy name (字, zì) and one or more art names (號, hào) besides his given name. After death, an honorable person may be awarded a posthumous name (諡號, shì hào), and an emperor is usually given a temple name (廟號, miào hào). Therefore, a great diversity of personal names can be found in the Ming and Qing novels, which are also used in other ancient Chinese literature. Annotation of such entities in great details enables Chinese literature works in different periods to be linked through extraction and reference of annotation tags. China's system of geographic units and administrative divisions also changes over time, which makes place name, organizational name, and official titles in the Ming and Qing novels different from those in modern Chinese. All of these features should be taken into consideration so that the specification can be applied to this kind of literature.

In general, the specification is designed based on the features of the Ming and Qing novels with reference to the existing word segmentation and annotation principles for modern Chinese and has been established iteratively as the tagging work progresses.

It should be pointed out that there was a related annotation work performed by Academia Sinica on Chinese classics [3]. The whole collection, called Tagged Corpus of Early Mandarin Chinese (hereinafter referred to as the Corpus), has included some classics of the Ming and Qing dynasties. Different from this work, the Corpus which focuses on grammatical tagging from the perspective of word category is fully parsed and annotated to facilitate filtering, searching, analyzing, and statistics of PoS. However, most of the named entities in the Corpus are put into one category and tagged as proper nouns only. In this case, semantic analysis would not be sufficient to link entities of literature in different historical periods which is part of our project objectives. In terms of words, the Corpus considers them more from the lexical side whereas we have more emphasis on semantic integrity. For example, the Corpus treats the segmentation of "桌/上"(zhuō/shàng, on the table) and "心/上" (xīn shàng, in one's heart) in the same way. But in our project, "心上" (xīn shàng, in one's heart) is not segmented because here "上" (shàng) is not a directional indicator.

---

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

[2] In *A Dream of Red Mansions*, "一心一計" (yī xīn yī jì, wholeheartedly) is found in chapters 6, 65, 69, 79, and 101, and "一心一意" (yī xīn yī yì, wholeheartedly) is only found in 98. Their meanings are exactly the same.

3 Principles of Word Segmentation

Based on the above analysis performed on the differences between the language of the Ming and Qing novels and modern Chinese, this work has formulated the basic segmentation rules applicable to the Ming and Qing novels.

First of all, a segmentation unit is defined as the smallest linguistic unit that "has specific semantic or grammatical functions", which is defined in GB13715, i. e., China's national segmentation standard for modern Chinese information processing [4]. The principles of word segmentation for the Ming and Qing novels are established from the perspectives of both semantics and syntax. From the perspective of semantics, the fundamental rule is to segment words into the smallest units without loss of semantic information, distortion, and ambiguity. A set of principles are also established from the perspective of syntax as complementary guidelines.

3.1 Word Segmentation Principles from the Perspective of Semantics

From the perspective of semantics, the basic segmentation principle is to segment character strings into the smallest units while there is no risk of meaning loss, misinterpretation, and ambiguity. For instance, "一/疾/而/終" (yī/jí/ér/zhōng, fell ill and died) is segmented into four units since each individual word has an independent semantic meaning or grammatical function. The following describes the main segmentation principles from the perspective of semantics accompanied by examples.

Handling of Functional Words. In the Ming and Qing novels, the following words are frequently used: "之" (zhī, usually used as a pronoun or a marker of subordination between nouns), "了" (le, a particle usually following a verb to indicate a completed action), "的" (de, a particle with flexible usages, such as attached to a pronoun or noun to indicate possession, to an adjective for description and emphasis, and to a verb for nominalization), "於" (yú, a preposition usually used to indicate time, place, or direction), "眾" (zhòng, used before nouns to indicate plural), "們" (men, used after nouns or pronouns to indicate plural), "只" (zhǐ, an adverb that means only, just, or simply), "被" (bèi, used as a marker of passive voice), "也" (yě, used as an adverb similar to "also", or used in the end of a sentence as a particle implying affirmation), "亦" (yì, an adverb that means "also"), "所" (suǒ, usually used in relative clauses and passives), "而" (ér, usually used as a conjunction to indicate parallel connection, temporal sequence, contrastive or concessive relations, etc.), "得" (dé, usually used as a particle following a verb to indicate the status of being able to), "時" (shí, commonly used as an adverb in the sense of "while" or "at that time"), "者" (zhě, usually used as a marker of nominalization). Since these words are commonly used in combination with different kinds of words to form various structures, they are regarded as independent segmentation units, for example, "用/了" (yòng/le, used) and "投降/者" (tóu xiáng/zhě, surrenders). However, fixed expressions are not segmented, for example, "我們" (wǒ men, we or us) and "作者" (zuò zhě, the author).

Handling of Demonstrative Pronouns. Demonstrative pronouns such as "這" (zhè, this), "那" (nà, that), "此" (cǐ, this), "某" (mǒu, a certain), etc. and words of inclusion or restriction such as "每" (měi, every), "各" (gè, each), "諸" (zhū, all), etc. are treated as segmentation units, for example, "這/石" (zhè shí, this stone). However, the lexical entries containing a pronoun referring to one or more unspecified beings are usually not segmented. This should be justified based on the context. For example, "閑步/至/此" (xián bù/zhì/cǐ, came out for a stroll and stopped here) and "事/已/至此" (shì/yǐ/zhì cǐ, indicating that something cannot be changed) are treated differently. In the former example, "此" (cǐ) refers to the place the character in the novel is arriving at, which is definite; while in the latter example, what "此" (cǐ) refers to is indefinite. So the same lexical entry is treated differently in different contexts.

Handling of Directional Words. The phrases in combination with words of locality are segmented if they indicate location or direction, for example, "門/前" (mén/qián, at the gate). Words of locality include "前" (qián, front), "後" (hòu, back), "左" (zuǒ, left), "右" (yòu, right), "上" (shàng, up), "下" (xià, down), "裏" (lǐ, in), "中" (zhōng, within), "內" (nèi, inside), "外" (wài, outside), "畔" (pàn, side), "旁" (páng, side), "邊" (biān, side or edge), etc. However, if the word groups have new meanings not associated with location or direction or there are no corresponding antonyms, they are not segmented, for example, "心下/乃/想" (xīn xià/nǎi/xiǎng, thinking to herself). The lexical entry "心下" (xīn xià, in one's heart) is not segmented because here "下" (xià) does not function as a directional indicator. In the Ming and Qing novels, "心上" (xīn shàng), "心下" (xīn xià), and "心中" (xīn zhōng) are all used, but they convey the same meaning "in one's heart" and neither "上" (shàng) nor "下" (xià) implies a direction.

Handling of Negation. The phrases in combination with negatives such as "不" (bù), "沒" (méi), "非" (fēi), "無" (wú), "勿" (wù), etc. are segmented, for example, "不/可" (bù/kě, should not). The context becomes a major factor for judging whether a phrase of this kind should be segmented. There are also some special cases:

- If a word group has new meaning or does not convey the negative meaning, it is not segmented. For example, "不但" (bù dàn, not only) is not segmented because it is not a negative.
- A fixed expression is not segmented. For example, although "不消" (bù xiāo, need not) is a negative, it is not segmented because it is used as a fixed expression.

Handling of Repetition. Duplicated words are not segmented, for example, "隱隱" (yǐn yǐn, half hidden). However, those inserted by a word such as "一" (yī, once) and "了" (le, a particle usually following a verb to indicate a completed action) are segmented, for example, "享/一/享" (xiǎng/yī/xiǎng, enjoy).

Handling of Words with Similar or Opposite Meanings. The phrases composed of two words with similar or opposite meanings are not segmented because they are usually used as fixed expressions, for example, "悲歡" (bēi huān, the joys and sorrows) and "抄錄" (chāo lù, copy). To help future study, a list of such lexical entries is appended to the specification document.

Handling of Suffixes. In Chinese, suffixes do not have independent semantic or grammatical functions, so they are not segmented from the words to which they are attached in this work. For example, "花兒" (huā ér, flower) is not segmented because it has the same meaning as "花" (huā, flower). The most commonly used suffixes in the Ming and Qing novels include "兒" (ér, mainly attached to nouns and verbs, used in dialects or informal situations), "子" (zǐ, mainly attached to nouns, sometimes also attached to verbs and adjectives for nominalization), "著" (zhe, mainly attached to verbs to indicate the unchanging state of an action), "些" (xiē, usually attached to verbs, adjectives, or pronouns to indicate indefinite amount or degree), and "然" (rán, mainly attached to adjectives and adverbs).

Handling of Idioms. Idioms, set phrases and the word groups with particular meanings different from the literal combinations of individual words are not segmented, for example "連二連三" (lián èr lián sān, in turn). It worth noting that there are some phrases used in Ming Qing novels, which may not be used in modern Chinese any longer, for example, "四下裏" (sì xià lǐ, everywhere).

3.2 Word Segmentation Principles from the Perspective of Syntax

Some rules are also formulated from the perspective of syntax as complementary guidelines. The following describes the main principles with examples.

Handling of Verb-Object Structure. The verb-object word groups are segmented if both the verb and the object have independent semantic functions, for example, "理/朝廷" (lǐ/cháo tíng, regulate the government). However, there are many special cases:

- 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, "嚼舌根" (jiáo shé gēn, describe sb. as foul-mouthed) is not segmented because it has a new meaning not associated with "嚼" (jiáo, chew) and "舌根" (shé gēn, tongue).
- The same lexical entry that has more than one grammatical function is treated differently in different contexts. For example, when "回書" (huí shū) functions as a predicate which means "to reply to one's letter", it is segmented into two units; however, when it is used as a noun with the meaning of "a letter in reply", it should not be segmented.

- The same lexical entry that has the same grammatical function may have different semantic meanings in different contexts. 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 literal sense of the individual words, that is, not associated with "下" (xià, down) and "馬" (mǎ, horse).

Handling of Adjective-Noun Structure. The adjective-noun word groups are segmented if both the adjective and the noun have independent semantic functions, for example, "奇/物" (qí/wù, something special).

Handling of Subject-Predicate and Predicate-Complement Structure. The word groups of this kind are usually segmented because the elements of these structures normally have independent semantic functions, for example, "吃/盡" (chī/jìn, finish off).

Handling of Combinations of Number, Quantifier, and Noun. The word groups of number, quantifier, and noun are segmented because the elements have independent semantic functions. These structures include "marker of ordinal numerals + number + quantifier", "number + quantifier + noun", "number + noun", etc. Here are two examples: "第/二/日" (dì/èr/rì, the next day) and "幾百/株/杏花" (jǐ bǎi/zhū/xìng huā, hundreds of apricot trees).

Handling of Adverb-Verb Structure. The adverb-verb word groups used as fixed expressions are not segmented, for example, "嚎哭" (háo kū, wail).

Handling of Combinations of Directional Verb and Directional Complement. In the Ming and Qing novels, many main verbs consist of a directional verb and a directional complement. Usually used as fixed expressions, they are not further segmented, for example, "上來" (shàng lái, come up) and "下去" (xià qù, go down). Directional verbs include "上" (shàng, up), "下" (xià, down), "過" (guò, cross over), "回" (huí, back), "進" (jìn, in), "出" (chū, out), "起" (qǐ, get up), "歸" (guī, go back), "到" (dào, get to), "走" (zǒu, walk), etc. Directional complements include "來" (lái, come), "去" (qù, go), "入" (rù, enter), etc.

4 Segmentation and Annotation of Named Entities

In this work, named entities most commonly used in the Ming and Qing novels are classified into six categories: personal name (人名), term of address (人物稱謂),

name of official position (官職) and title of nobility or honour (爵位、封號), place name (地名), building name (建築名), and organizational name (組織名).

In the Ming and Qing novels, there are many compound named entities. For instance, various terms of address are generated by different combinations of any form of a person's name with a title indicating the person's rank or position. Any personal name, term of address, or title may also be used as a part of a place name or a building name. That is why a variety of compound named entities are formed. This study seeks to establish a set of annotation rules for different kinds of named entities to ensure that they can be segmented and tagged in a consistent and flexible way. The establishment of these rules is based on researches on Chinese literature, analysis and statistics in combination with experience in actual annotation. In general, the approach of Peking University for named entity tagging [1] is followed: 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 stated in this paper.

To facilitate future research and application in literature study, this work also distinguishes real persons and places from fictional ones. Even though persons and places in literature works are usually fictional, there are many references to real historical figures and places in Chinese classics due to the rich history of China. For example, in the novel *Romance of the Three Kingdoms*, there are many references to real persons and places which can be found in the official records. To make a distinction and enable easier identification for the study of Chinese classics, personal names, terms of address, place names, and building names are further classified into four categories as described below:

- Real entities: labeled with "#", for example, "蘇軾/nr3#" (Sū Shì/nr3#, one of the major poets of the Song dynasty). In this work, the entities recorded in *The Twenty-Four Histories* (《二十四史》) are considered as real ones.
- Mythical entities: labeled with "*", for example, "灌愁海/ns3*" (Guàn Chóu Hǎi/ns3*, the Sea of Brimming Grief).
- Fictional entities cited from other literature works: labeled with "&", for example, "紅娘/nr6&" (Hóng Niáng/nr6&, a maid in *Romance of the West Chamber* 《西廂記》) which is mentioned in a dialogue of *A Dream of Red Mansions*.
- Fictional entities in the novel being processed: this is default with no special symbol required.

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

4.1 Personal Names

Various forms of names in the Ming and Qing novels are categorized into six types: surname, given name, surname + given name, courtesy name, surname + courtesy

name, and alternative name. This section describes the principles for tagging personal names in details.

Surname. The most common surname contains only one character, but there are also multi-character surnames in China. It is unnecessary to segment a surname no matter how many characters it contains. Any surname referring to a specific person is tagged with "/nr1", for example, "薛/nr1 林/nr1 二/人" (Xuē/nr1 and Lín/nr1, referring to Xue Baocai and Lin Daiyu, two main characters in *A Dream of Red Mansions*).

Given Name. It is unnecessary to segment a given name. Any given name referring to a specific person is tagged with "/nr2", for example, "黛玉/nr2" (Dàiyù/nr2).

Surname + Given Name. A complete personal name is composed of a surname name plus a given name. Any "surname + given name" referring to a specific person is tagged with "/nr3". There are different cases:

- If the surname contains only one character, it is the default case and there is no need to separate it from the given name because the computer will regard the first character as the surname in this case, for example, "林黛玉/nr3" (Lín Dàiyù/nr3).
- If the combination of "surname + given name" contains a multiple-character surname, the surname is separated from the given name with "//", for example, "司馬//相如/nr3#" (Sīmǎ//Xiàngrú/nr3#, an official of the Western Han Dynasty well-known for his prose poems).
- Sometimes, when mentioning a woman, people may add her husband's surname before her own surname. If a compound name contains more than one surname, the surnames are separated with "//" and they are also separated from the given name with "//". In this way, different kinds of combinations can be treated consistently and flexibly, such as "two-character surname//given name", "two-character surname//one-character surname//given name", and "one-character surname//two-character surname//given name". Thus, both surnames and given names can be recognized by the computer easily.

Courtesy Name (字, Zì). In traditional Chinese culture, an adult male usually selects or acquires from other people a courtesy name as a symbol of adulthood and respect. As another form of a given name, it commonly consists of one or two characters. It is unnecessary to further segment a courtesy name and the whole courtesy name is tagged with "/nr4", for example, "孔明/nr4#" (Kǒngmíng/nr4#, the courtesy name of Zhuge Liang, a famous strategist during the Three Kingdoms period of Chinese history).

Surname + Courtesy Name. Any "surname + courtesy name" referring to a specific person is tagged with "/nr5" with other rules similar to that of "surname + given

name", for example, "諸葛//孔明/nr5#" (Zhūgě//Kǒngmíng/nr5#, the combination of surname and courtesy name of Zhuge Liang).

Alternative Name. Besides given name and courtesy name, people may have some other alternative names, including milk name, nickname, pen name, art name (號, hào, an alternative courtesy name most commonly three or four characters in length), posthumous name (諡號, shì hào, a honorary name selected after a person's death), temple name of an emperor (廟號, miào hào), etc. All these alternative names are put into one category. Any alternative name referring to a specific person is tagged with "/nr6", for example, "顰兒/nr6" (Pín ér/nr6, a nickname of Lin Daiyu). There are some special cases:

- Since most names of foreign nations and races in the novels are translated or transliterated from foreign languages, they are treated as alternative names if the surname and the given name cannot be identified, for example, "金環三結/nr6" (Jīn Huán Sān Jié/nr6, one of the chiefs of the tribesmen).
- A person's name may be changed for different reasons. For example, in ancient China, the names of the emperors, elders, and people of higher rank are regarded as taboos, so a person's name may have to be changed. Whatever the reason is, the new name is regarded as an alternative name.
- If a compound alternative name contains a surname with only one character, it is unnecessary to separate it from the following alternative name. If a compound alternative name contains one surname with two or more characters, or it contains more than one surname, the principle is the same as that for "surname + given name".

4.2 Terms of Address

A variety of address terms can be found in the Ming and Qing novels. If they are not tagged in the corpus, word sense ambiguity may be caused. For example, 公 (Gōng) has many meanings when used as a common adjective, such as public, fair, etc. In the Ming and Qing novels, it is also used as a term of address in respectful term or as the title duke. If the named entities are not properly tagged, this kind of information cannot be identified and extracted efficiently. Terms of address in the Ming and Qing novels fall into two categories. In the first category, it is used alone without being combined with personal names whereas in the second category, it is used in combination with other names given in Section 4.1.

Terms of Address Used Alone. A term of address may indicate a person's gender, marital status, kinship, social class, occupation, religious belief, etc. Any address term of this kind referring to a specific person is tagged with "/na2", for example, "老爺/na2 說/了" (lǎo yé/na2 shuō/le, the master says). Terms of address are segmented according to the word segmentation principles stated earlier. Square brackets are used to enclose those consisting of more than one segmentation unit, for example, "[二/小姐]/na2" ([èr/xiǎo jiě]/na2, the Second Young Lady). Only the term of address

referring to a particular person that can be identified from the context is tagged. For example, the following terms of address are not tagged: "一/個/小/丫頭/扶/了" (yī/gè/xiǎo/yā tóu/fú/le, leaning on a young maid's arm), "老爺/們" (lǎo yé/men, the masters).

Terms of Address Combined with Any Form of Name or Title. A term of address may be combined with a surname, given name, title, etc. to form a compound one. Any address term of this kind referring to a specific person is tagged with "/na1". The whole compound address term is enclosed in square brackets, in which the units are segmented and tagged according to the principles stated in this paper, for example, "[政/nr2 老爺]/na1" ([Zhèng/nr2 lǎo yé]/na1, Lord Zheng). This principle ensures that all kinds of complicated compound address terms are treated in a consistent way. If a compound address term contains more than one surname, the surnames are separated with "//" and they are also separated from the term of address with "/", for example, "[張//王/nr1 氏]/na1" ([Zhāng//Wáng/nr1 shì]/na1, in which Wáng is the surname of a woman, Zhāng is her husband's surname, and here shì is used to address a married woman).

4.3 Names of Official Position and Titles of Nobility or Honour

A name of official position is tagged with "/nu1". There are also many compound ones used together with an organization or a place. In this case, they are also treated in the same way as other compound named entities, for example, "[太醫院/nt 正堂]/nu1" ([Tài Yī Yuàn/nt zhèng táng]/nu1, the director of the Academy of Imperial Physicians) where "/nt" refers to an organization.

The system of nobility and honorific titles is an important feature of Chinese culture in the imperial age. Of course titles of nobility vary from dynasty to dynasty, but the most common five ones are used almost throughout China's whole imperial history: 公 (Gōng , duke), 侯 (Hóu, marquis), 伯 (Bó, earl), 子 (Zǐ, viscount), and 男 (Nán, baron). A title of nobility or honour may be granted to members of the imperial house and their blood relatives and in-laws. It may also be bestowed on persons of high merits or heroes who have made great contributions. In addition, there is also a well-developed system of titles for female members of the aristocracy. All of these noble and honorific titles are put into one category and tagged with "/nu2", for example, "[保齡/侯]/nu2", ([Bǎolíng/Hóu]/nu2, Marquis of Baoling). Sometimes, the name of the land granted to a noble person is attached to the title of nobility or honour. In this case, it is also treated in the same way as other compound named entities, for example, "[烏程/ns2# 侯]/nu2" ([Wūchéng/ns2# Hóu]/nu2, Marquis of Wucheng).

4.4 Place Names

Places are generally classified into four categories:

- **Country:** A country name is tagged with "/ns1", for example, "暹羅國/ns1#" (Xiān Luó Guó/ns1#, Siam, the former name of Thailand).

- **Prefecture, city, county, township, village, street, road, etc.:** This kind of place name is tagged with "/ns2", for example, "金陵/ns2#" (Jīnlíng/ns2#, the present-day Nánjīng).
- **Mountain, grassland, river, lake, sea, island, etc.:** This kind of place name is tagged with "/ns3", for example,"灌愁海/ns3*" (Guàn Chóu Hǎi/ns3*, the Sea of Brimming Grief).
- **Other places:** Other place names are tagged with "/ns4", for example, "虎牢關/ns4#" (Hǔláo Guān/ns4#, Hulao Pass, a mountain pass which is the site of many historical battles).

4.5 Building Names

Buildings in the Ming and Qing novels are broadly divided into three categories:

- **Palace, mansion, official residence, private garden, pavilion, etc.:** This kind of building name is tagged with "/nv1", for example, "大觀園/nv1" (Dà Guān Yuán/nv1, the Grand View Garden).
- **Temple, pawnshop, restaurant, teahouse, and some other public spaces:** This kind of building name is tagged with "/nv2", for example, "葫蘆廟/nv2" (Hú Lu Miào/nv2, the Gourd Temple).
- **Other buildings:** Other building names are tagged with "/nv3", for example, "翠煙橋/nv3" (Cuì Yān Qiáo/nv3, the Green Mist Bridge).

Sometimes, a building name is combined with a surname, alternative name, title etc. In this case, it is treated in the same way as other compound named entities: the whole compound building name is enclosed in square brackets, in which the units are segmented and tagged accordingly. The following gives two typical examples: "[榮/nu2 府]/nv1" ([Róng/nu2 fǔ]/nv1, the Rong Mansion, which is the residence of the Duke of Rongguo and his descendants), "[[忠靖/侯]/nu2 史/nr1 府]/nv1" ([[Zhōngjìng/Hóu]/nu2 Shǐ/nr1 fǔ]/nv1, the residence of Marquis of Zhongjing whose surname is Shǐ).

4.6 Organizational Names

The name of a particular organization is tagged with "/nt", for example, "太醫院/nt" (Tài Yī Yuàn/nt, Academy of Imperial Physicians). The compound organizational name containing more than one segmentation unit is treated in the same way as other compound named entities.

5 Quality Assurance and Annotation Evaluation

To ensure the quality of the annotated corpus, the project takes the approach of computer-aided annotation rather than using purely computerized or purely manual approach. The process of segmentation and tagging mainly consists of three stages:

system training and automatic processing, manual annotation and review, and post-processing.

In the first stage, literature dictionaries of some classics and a small amount of named entities which are manually identified are integrated into the existing segmentor/tagger [5]. In this way, the segmentor/tagger, which was originally developed to process modern Chinese, is trained to process the Ming and Qing novels. Then the software is used to segment and tag the text automatically. In the second stage, named entities are tagged and the whole text is manually reviewed by one annotator at least twice in strict accordance with the specification stated in this paper. In the third stage, a software tool is used to check inconsistency. Even though this process works in stages, iterations may be required from time to time as new cases in manual review and consistency check may make it necessary to slightly revise the specification and update the dictionaries.

To evaluate the quality of the output produced by the segmentor/tagger, ten chapters, 67,181 characters in length, were randomly selected from *A Dream of Red Mansions* for evaluation. The output compared to the result achieved after the 3rd stage shows that the automatically processed text can achieve 90.14% in precision and 94.48% in recall without human intervention. To evaluate the quality of the result in the 2nd stage, ten paragraphs were randomly selected from *A Dream of Red Mansions*, and the differences generated during the cross-check were discussed with the annotator to set an agreed golden answer. The results show that the precision is 99.44%. About 0.31% of all the errors are caused by different interpretations of different people on some lexical entries or on the specification. Once the specification is refined, these errors can be further reduced. The rest of errors, about 0.25% of the total, are human errors which are difficult to be eliminated completely. According to the specification, some lexical entries in the dictionaries integrated into the system should be further segmented manually. During manual review, it is possible to ignore a few entries. If necessary, the dictionaries will be updated accordingly. After correction of the errors, the specification document and the dictionaries will also be updated to improve the performance of automatic segmentation and annotation. The target is to ensure that the final corpus can reach a precision of 99.5%. In view of the quality assurance measures taken throughout the whole process, this goal is achievable, which is proven by the previous test data.

6 Conclusion

This study presents the specification for segmentation and annotation of Chinese novels in the Ming and Qing dynasties from the perspectives of semantics and syntax based on the standards for modern Chinese segmentation widely accepted in Mainland China [1] and in Taiwan [2]. This specification is designed in consideration of the unique characteristics of the novels in that period compared to modern Chinese, including frequent use of single-character words, the chapter-by-chapter style, the combinations of compound named entities, idiomatic expressions, etc. In terms of segmentation, the fundamental principle is to segment text strings into the smallest

linguistic units with independent semantic or grammatical functions while there is no risk of meaning loss, distortion, misinterpretation, and ambiguity. Different types of named entities, especially the compound ones flexibly grouped by various elements, are segmented and tagged in a consistent and flexible way. In the end, this paper also shows the quality control process, the measures taken in practical work, and a preliminary evaluation of the annotation quality in different stages of this work. The annotated corpus built in accordance with the specification can be used in different fields of study such as linguistics, literature, history, teaching of Chinese, and even information technology. The results of this study are expected to lay a foundation for computer-aided semantic analysis and named entity tagging of Chinese literature works in the Ming and Qing dynasties and more ancient times.

Acknowledgments. This work 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. Yu, S.W., Duan, H.M., Zhu, X.F., Swen, B., Chang, B.B.: Specification for Corpus Processing at Peking University: Word Segmentation, POS Tagging and Phonetic Notation. Journal of Chinese Language and Computing 13(2), 121–158 (2003) (in Chinese)
2. Segmentation Principle for Chinese Language Processing (CNS14366). National Bureau of Standard, Taiwan (1999) (in Chinese)
3. Wei, P.C., Thompson, P.M., Liu, C.H., Huang, C.R., Sun, C.F.: Historical Corpora for Synchronic and Diachronic Linguistics Studies. International Journal of Computational Linguistics & Chinese Language Processing 2(1), 131–145 (1997) (in Chinese)
4. Liu, Y., Tan, Q., Shen, X.K.: Segmentation Standard for Modern Chinese Information Processing and Automatic Segmentation Methodology (GB13715). Qinghua University Press, Beijing (1994) (in Chinese)
5. Lu, Q., Chan, S.T., Xu, R.F., Chiu, T.S., Li, B.L., Yu, S.W.: A Unicode based Adaptive Segmentor. Journal of Chinese Language and Computing 14(3), 221–234 (2004)