A Study on Personal Attributes Extraction Based on the Combination of Sentences Classifications and Rules

Nan-chang Cheng
Institute of Automation Chinese Academy of Sciences
Beijing, china
nanchang.cheng@nlpr.ia.ac.cn

Cheng-qing Zong
Institute of Automation Chinese Academy of Sciences
Beijing, china
cqzong@nlpr.ia.ac.cn

Min Hou
Communication University of China
Beijing, china
houmin@cuc.edu.cn

Yong-lin Teng
Communication University of China
Beijing, china
tengyonglin@cuc.edu.cn

Abstract

Personal attributes extraction plays a significant role in information mining, event tracing and personal name disambiguation. It mainly involves two problems, attribute recognition and decision making on whether this attribute belongs to the extracted person. Personal attributes generally involve named entities, which are recognized mainly by adjusting word segmentation software. As for those which cannot be recognized by word segmentation, the combination of feature words and rules can be used for their recognition. The combination of sentences classifications and rules is employed for attribute ownership decision. At first, all the sentences in the document are classified into those with attribute words and those without, with the latter omitted. The former are then classified into description sentences with one person and description sentences with more persons, according to the criterion that whether there are more than one person described in the sentence. According to statistics of description sentences with one person, anaphora resolution is not necessary, which reduces recognition errors from anaphora resolution failures. Minimum slicing is used for description sentences with more persons, and attribute ownership decision is made within the minimum language segment with the co-occurrence of both the person and the attribute. This method achieves 0.507388780 and 0.489505010 respectively in the lenient evaluation results and the strict evaluation results of SF_Value in CIPS-SIGHAN2014[1] Bakeoff, which turns out to be the best. The fact has shown that the method is effective.

1 Introduction

Attribute, characterized by its objectivity, is inherent in things(Zhuang, 2000). Personal attribute extraction aims at automatically extracting in unstructured texts specific attributes associated with the personal name, such as the character entity's date of birth, work units, spouses, children, education, title, etc. This plays a significant role in information mining, event tracing and personal name disambiguation. International TAC KBP has been conducted since 2009 (Bikel et al., 2009; McNamee et al., 2009), and CIPS-SIGHAN2014 has referred to and revised its Slot Filling tasks to design personal attribute extraction tasks in Chinese. There are six groups participating this bakeoff.

Personal attribute extraction mainly involves two problems, attribute recognition and decision making on whether this attribute belongs to the extracted person, and the latter can be called attribute ownership decision. Personal attributes are generally named entities, such as personal names, place names, organization names, temporal nouns, so named entity recognition technology is needed in attribute recognition. Although named entity recognition is one difficulty in natural language processing, there are plenty of experiences and methods we can draw upon as 30 years has witnessed its research since the introduction of Chinese word segmentation, as in (Sun et al., 1995; Zhao et al., 1999; Liao et al., 2004; Yu et al., 2006; Ye et al., 2007). Therefore, this paper focuses upon attribute ownership decision after a brief introduction to personal attribute extraction, since the former is more complicated with anaphora resolution and attribute ownership decision among more persons. Some of bakeoff papers

[1] http://www.cipsc.org.cn/clp2014/webpage/en/home_en.htm

Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 192–201,
Wuhan, China, 20-21 October 2014

regarding filling slot have noticed these problems, as in (Bikel et al., 2009; Burman et al., 2012). In this paper, we propose attribute ownership decision through the combination of sentences classifications and rules in accordance with natural language features and the task requirements of our bakeoff. This method has achieved good results in the evaluation. The rest of the paper is organized as: Section 2 introduces main ideas, Section 3 presents the methods of personal attribute recognition, Section 4 emphasizes on and discusses the methods of personal attribute ownership decision, Section 5 is experimental results and Section 6 is conclusion.

2 Main ideas

Attribute recognition is mainly named entity recognition, which is attempted to be settled in word segmentation in our study. According to attribute recognition task requirements, the word segmentation software used in this study has been adjusted so that it can recognize most named entities. As for those which cannot be recognized by the software, the method of feature words together with rules has been employed. After attribute recognition, all the sentences in the document are classified into those with attribute words and those without, with the latter omitted. Therefore, attribute ownership decision is merely conducted to the sentences marked with attribute words.

Now that the anaphora of personal pronouns are widely used in most sentences, attribute ownership decision involves anaphora resolution, which means the determination of the antecedent of the anaphor(Wang, 2005). Anaphora resolution appears to be difficult in Chinese, far from being settled completely satisfactorily(Wang, 2002; Wang 2005). In order to decrease the reliance on anaphora resolution, we have studied the tested documents and found that the described person in most of them is the extracted character. When most sentences in a document describe the extracted person, it is not necessary to employ anaphora resolution. Anaphora resolution or some other methods are needed to find the attribute of the extracted person only for those sentences with more persons. In a small number of documents, there is only one extracted person within the whole text, such as “马伟明_T1.xml” and “白志东_T1.xml”. As such, in attribute ownership decision, it should be determined whether there are more than two persons described in the sentence. In this way, the sentences marked with attribute words in the document will be classified as description sentences with one person and description sentences with more persons through some methods, which would decrease the reliance on anaphora resolution and so greatly improve decision precision by decreasing the recognition errors from anaphora resolution failures. The challenge here is how to determine those sentences with more persons, which will be expounded later.

3 Personal attribute recognition

Personal attribute recognition involves two jobs. One is to adjust word segmentation software in order to achieve full recognition of various types of named entities, and the other is to annotate feature words to ensure exact decision of attribute identity of some named entities.

3.1 Adjusting word segmentation software

Named entity recognition is mainly completed in word segmentation. The word segmentation software used is CUCBst, a dictionary and rule based software developed by Broadcasting Media Center, Communications University of China. The adjustment includes: adjustmenting tagging, adding words, and adjusting rules.

3.1.1 Adjustmenting Tagging

First, some tags are adjusted in the dictionary. Take some words associated with titles as an example. In the dictionary, there are items such as “程序员(programmer) n”, “雕刻师(sculptor) n”, “董事长(president) n”, “发明家(inventor) n” and “检察长(chief-prosecutor) n”. The tag of “n” within is adjusted to be “tt”. For instance:

Example Sentence 1: 可见魏冉这位封建社会地主阶级的政治家，在完成秦王朝统一中国的事业中所起的作用。

Translation: As a feudal politician of landlord class, in the cause of uniting China by Qin dynasty, Wei Ran's role is clearly demonstrated.

Its tagged version is:

可见/c 魏冉/nr 这位/r 封建社会/ln 地主/n 阶

级/n 的/u 政治家/tt，/w 在/d 完成/v 秦王朝/t 统一/a 中国/gj 的/u 事业/n 中/f 所/u 起/v 的/u 作用/n 。/w

Through the tagging adjustment, it is easy to recognize the title of the extracted person “魏冉(Wei Ran)” is “政治家(statesman)”. We also adjust the tagging of death reasons(sw), nations(gj), provincial cities(sh), cities and towns(sx). In addition, some feature words are annotated. For example, the feature words associated with character birth such as “生于(be born)”, “出生(be born)” and “诞生(be born)” are annotated as “bir”.

3.1.2 Adding words

There are two stages in adding words:

Stage One is to collect and sort dictionaries in system development, adding names such as titles, nations and places to the segmentation dictionary.

Stage Two is to add OOV words to the segmentation dictionary in evaluation period by implementing new words automatic recognized in evaluation corpus with manual intervention. It should be pointed out that some certain noun phrase is regarded as one word and then kept in the dictionary. These noun phrases are mainly organization titles, nicknames and titles such as “北平研究院物理研究所(Institute of Physics of Peking Academy of Sciences)”, “罗彻斯特储蓄银行(Rochester Bank)”, “橙县小姐(Miss Orange County)” and “名誉理事长(Honorary chairman)”.

3.1.3 Adjusting rules

CUCBst segmentation system is characterized by coarse-grained segmentation and fine-grained segmentation, which is implemented by rules. We adjust some merging rules so that they can achieve better attribution recognition. For example:

Example Sentence 2: 斯托曼 1953 年出生于美国纽约曼哈顿地区的犹太人家庭。

Translation: Stallman was born of a Jewish family in Manhattan, New York, in 1953.

Its segmented version before the rule adjustment is:

coarse-grained segmentation：斯托曼/nr 1953 年/t 出生/v 于/p 美国纽约曼哈顿地区/ns 的/u 犹太人/n 家庭/n

fine-grained segmentation：斯托曼/nr 1953 年/t 出生/v 于/p 美国/ns 纽约/ns 曼哈顿/ns 地区/n 的/u 犹太人/n 家庭/n

In the coarse-grained segmentation version, “美国纽约曼哈顿地区”, which includes two personal attributes in accordance with evaluation outline, country of birth and city of birth, is merged together. Further analyses and processes are needed for correct recognition. In the fine-grained segmentation version, “美国纽约曼哈顿地区” is divided into 4 words as “美国/ns 纽约/ns 曼哈顿/ns 地区/n”, in which country of birth is correctly segmented. However, city of birth needs further processes by merging the following three words. Example Sentence 2’s segmented version after the rule adjustment is:

斯托曼/nr 1953 年/t 出生/v 于/p 美国/gj 纽约曼哈顿地区/ns 的/u 犹太人/n 家庭/n

In this version, “美国纽约曼哈顿地区” is segmented into 2 words as “美国/gj 纽约曼哈顿地区/ns”, which are country of birth and city of birth respectively. This makes the recognition and extraction of related attributes convenient.

3.2 Finding nearest named entity through the feature word

Although some specific tagging aimed for named entities and some personal attributes is conducted in word segmentation, it should be noted that not all tagged named entities are personal attributes. For example, 1998 is not always a person’s date of birth, since it could be the date for an event or something else. Therefore, it is necessary to decide personal attribute through the feature word, and find nearest named entity through the feature word within the sentence. Take the example of time of birth:

Example Sentence 3: 张幼仪/nr 生于/bir 1900 年/t ，/w 比/p 徐志摩/nr 小/a 4/m 岁/q 。/w

Translation: Zhang Youyi was born in 1900, and she was four years younger than Xu Zhimo.

Example Sentence 4: 鲁桂珍/nr 1904 年/t 生于/bir 南京/ns

Translation: In 1904, Lu Guizhen was born in Nanjing.

When segmented, “生于(be born)” is tagged as “bir”, which means the word is a feature word associated with a person’s birth. When there is “bir” in a sentence, the system will iterate before and after this feature word to find the nearest time noun, as in Example Sentence 3, 1900 is after the feature word and in Sentence 4, 1904 is

before the feature word.

4 Deciding whether the attribute belongs to the extracted person

In this section, we first classify the sentences in two levels in order to decide the attribute ownership in the classified sentence. As for the description sentence with one person, decide whether the character is the extracted object. If not, just omit the sentence. As for the description sentence with more persons, decide personal attribute ownership by extracting the personal attribute within the minimum language segment with the co-occurrence of both the person and the attribute.

4.1 Sentence classification

Sentence classification involves two levels. First, the sentences are classified into sentences with or without attribute marks. Then, classify the sentences with attribute marks into those with one person and those with more persons.

4.1.1 Classifying all the sentences into two types

All the attributes and feature words are marked in word segmentation. In terms of these marks, all the sentences are classified into two types. Those without attribute marks will be directly omitted, whereas those with attribute marks will be kept for further processing.

4.1.2 Classifying the sentences with attribute marks into two types

The sentences with attribute marks are classified into those with one described person and those with 2 or more than 2 described persons. Character recognition is significant in this step. The forms to recognize characters include personal names, only surnames or first names, personal pronouns, zero form and kinship titles, in which personal names and kinship titles can be either antecedent or anaphora, the rest three can only be anaphora.

(1) personal names

Personal names are the most important feature to detect characters. For example:

Example Sentence 5. 1973 年 7 月 19 日，冯白驹在北京逝世。

Translation: On July 19, 1973, Feng Baiju passed away in Beijing.

Example Sentence 6. 次年 1 月，王文明病逝，冯白驹继任中共琼崖特委书记。

Translation: In January of the next year, Wang Wenming passed away, and Feng Baiju take Wang' place to be the Special Secretary of CPC in Qiongya.

Here, the number of personal names in the sentence will decide whether the sentence is the one with one described person. Example Sentence 5 is the sentence with one described person, for there is one personal name “冯白驹” within, whereas Example Sentence 6 is the sentence with more described persons, for there are two personal names within, “王文明” and “冯白驹”.

(2) only surnames or first names

As for non-Chinese names, the whole name is used first and then generally the surname is used for anaphora. For example:

Example Sentence 7. 莫奈 1840 年 11 月 14 日出生于法国巴黎 45 街拉菲特第九郡，是阿道夫和路易斯的第二个儿子。（克劳德·莫奈）

Translation: Monet was born on November 14, 1840, in 45 Street, the 9th canton of Lafayette, Paris, France; and he was the second son of Adolf and Louis. (Claude Monet)

When using the surname would be confusing, first names will be used, as in the introduction to the twin brothers, “Mike Bryan” and “Bob Bryan”, in Example Sentence 8.

例句 8. 等到鲍勃和迈克开始真正对网球产生了浓厚兴趣，也拿起球拍开始了网球生涯后，布莱恩夫妇又给他们订了个规矩：在 17 岁之前，这对双胞胎都不可以在比赛中对抗。

Translation: Until Bob and Mike really grew strong interests in tennis and began their tennis career, Bryans set up a catch for them. Before 17 years old, the twins were not permitted to compete in game.

Generally speaking, the whole name is used for the Chinese name. However, only surnames or first names could be used. For example:

Example Sentence 9. 七七事变后，日本人邀请他组建“中日友好协会”，梁意识到，要想不当汉奸，必须立即离开北平。（梁思成）

Translation: After Marco Polo Bridge Incident of 7th July 1937, the Japanese invited him to organize the "China-Japan Friendship

Association", Liang realized that he had to leave Peking immediately; otherwise he would be forced to become a traitor. (Liang Sichneg)

Example Sentence 10. 我与泽涵兄交往多了，与他的家人都处得很熟。（江泽涵）

Translation: After frequent contacts with Bro Zehan, I got well acquainted with his families. (Jiang Zehan)

When only surnames or first names are used, it is a little difficult to recognize them. Once recognized, it is as easy to decide whether there is one person or there are more persons in the sentence, as in the case of personal names.

(3) personal pronouns

Anaphora means that another component is used to refer to the prior component in order to avoid its repeat in the text(Xu 2003). There are three forms of anaphora, zero anaphora, pronominal anaphora and NP anaphora(Chen, 1987). In the personal attribute extraction, personal pronouns are anaphora with obvious forms and are used as one of the features to detect characters. For example:

Example Sentence 11. 江泽涵是中国数学会的创始人之一，从 1935 年该会成立时起，他就是副理事长。（江泽涵）

Translation: Jiang Zehan, one of the founders of the Chinese Mathematics Society, has been the vice chairman since the association was founded in 1935. (Jiang Zehan)

When the character is detected, a single personal pronoun (such as he, she, you and I) used in one sentence, even with several occurrences, will be regarded as only one person, for in one sentence, it is rare to use the same single personal pronoun to refer to different persons.

Generally the sentence with plural personal pronouns includes more persons. For example:

Example Sentence 12. 李约瑟一如既往忠于他的爱妻："执子之手、与子偕老，"直到 1987 年德萝西 91 岁时去世，他们夫妇共同生活了整整 64 年。

Translation: Joseph Needham was always loyal to his beloved wife, just as the famous Chinese saying goes, "Holding your hand, lead our merry life till old". Until De Luoxi left at the age of 91 in 1987, the couple had lived together for a full 64 years.

(4) kinship titles

When the extracted person is introduced, some other related persons will be mentioned. Relatives, such as parents, the wife and brothers, are often mentioned. Besides, some other connections may also be mentioned, such as teachers, friends and leaders. The kinship titles have obvious form features and can be used for detecting characters in the sentence. For example:

Example Sentence 13. 布兰切特的降生充满了浪漫色彩，爸爸是美国前海军军官，军舰在澳洲墨尔本停靠时，与布兰切特的母亲相识。（凯特·布兰切特）

Translation: Blanchett birth is full of romance. His father, a former US Navy officer, met Blanchett mother when the warship docked in Melbourne, Australia. (Cate Blanchett)

In Example Sentence 13, there are three persons, "Blanchett", "father" and "mother".

In addition, we also find that when some attributes of the extracted person's teacher, student, friend or leader are described, this person's name will appear. However, when a teacher, a student or a professor is used in a general sense, he or she has little thing to do with attribute extraction, so he or she will not be regarded as a character. For example:

Example Sentence 14. 但法伊弗却透露，自己上高中的时候很不受欢迎，"我那时很高，很笨拙，老师曾经在我的成绩单上写过'米歇尔是班里个子最大的女孩'"。

Translation: But Pfeiffer has revealed that she was very unpopular in high school, "At that time, I am very tall but somehow clumsy, and my teacher once wrote on my report card 'Michelle is the tallest Girl in class'".

Example Sentence 15. 梅耶的死让很多人震惊，他的同事和学生认为他是一个非常有才华的科学家和教师。

Translation: Meyer's death shocked a lot of people, both his colleagues and students believed that he was a very talented scientist and teacher.

Example Sentence 16. 需要提出的是，卡罗瑟斯的学生 Paul J. Flory (1910-1985)，在总结研究卡罗瑟斯的基础上，出版了影响整个世界的《高分子化学原理》一书，该书依然是今天高分子领域主要的理论基础。

Translation: I must point out that Carothers' student Paul Flory (1910-1985), on the basis of summarizing research on Carothers, published "Principles of Polymer Chemistry", which shook

the whole globe. The book is still the bible-like theoretical basis of today's realm of polymer.

“老师(the teacher)” in Example Sentence 14, “同事、学生(colleagues, students)” in Example Sentence 15 are used in a general sense, so both sentences are ones with one person. Instead, Example Sentence 16 makes clear the date of birth, date of death, and some other information, concerning Carothers’ student, Paul J. Flory(with a specific name for the student), so the sentence is one with more persons.

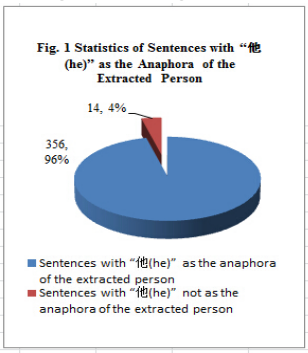
4.2 Attribute ownership decision

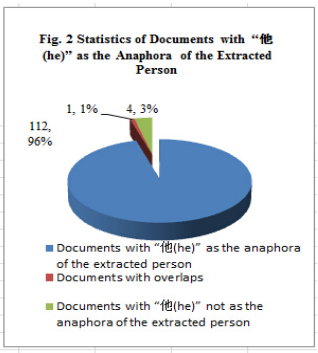
By employing the above mentioned character recognition features to classify the sentences, we get two sentence sets, the description sentences with one person (including zero anaphora) and the description sentences with more persons.

4.2.1 The description sentences with one person

(1) affirming the extracted person

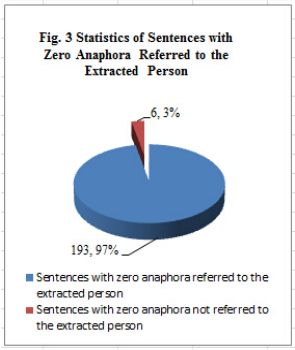
As for the sentences with personal names, including with only first names or surnames, the extracted persons’ names, including first names or surnames, are used for the match. The difficulty lies in the sentences with personal pronouns and zero form. As mentioned above, most documents in the testing texts mainly describe extracted persons, thus when the description sentences with one person involve personal pronouns and zero form, it can be hypothesized that extracted persons are directly used as described persons. In order to test this hypothesis, we study the use of the third singular personal pronoun “他(he)” in all the sentences. Through automated recognition, we obtain 369 sentences with one person which have “他(he)”. Then we identify all the sentences to see whether “他(he)” is the anaphora of the extracted person. Fig. 1 and Fig. 2 show the results.

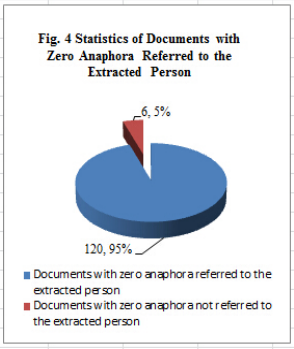




As illustrated in Fig. 1 and Fig. 2, 356 sentences, in 112 documents, with “他(he)” as the anaphora of the extracted person, account for 96 percent of all the sentences, whereas 14 sentences, in 5 documents, with “他(he)” not as the anaphora of the extracted person, account for only 4 percent of all the sentences. We study these 5 documents and find that the chiefly described person is not the extracted person in 3 documents, which are “鲁桂珍_T2.xml”, “鲁桂珍_T3.xml” and “陈济棠_T3.xml”. In “鲁桂珍_T2.xml” and “鲁桂珍_T3.xml”, the chiefly described person is 鲁桂珍’s husband, 李约瑟, not the extracted person, 鲁桂珍. In “陈济棠_T3.xml”, he chiefly described person is 陈济棠’s son, 陈树柏, not the extracted person, 陈济棠. In this document, there are 5 sentences with one person which have “他(he)”. There are 4 sentences with “他(he)” not as the anaphora of the extracted person, while there is only one sentences with “他(he)” as the anaphora of the extracted person. Thus we call this document as one with overlaps. The other two documents are “马伟明_T3.xml” and “白志东_T3.xml” respectively. Although the chiefly described person is the extracted person in both documents, the narrative perspective is first-person perspective.

In addition, we also perform statistical analysis of the use of zero form. As there are a number of zero anaphora, 193 sentences with zero anaphora are randomly chosen from 126 documents. Then we identify all these sentences to see whether there is the anaphora of the extracted person. The results are shown in Fig. 3 and Fig. 4.





As illustrated in Fig. 3 and Fig. 4, zero anaphora shares similar use with the anaphora of the third singular pronoun “他(he)”. By analyzing the documents with zero anaphora not referred to the extracted person, we find that the chiefly described person is not the extracted person.

However, there is no first-person perspective, which is quite different from the case of the third singular pronoun “他(he)”.

The data above demonstrate that our hypotheses are in line with reality. If we have had classified the documents in terms of some features such as the chiefly described person and narrative perspectives and then classified the sentences in documents, we would have achieved better results.

(2) attribute extraction

The extracted character in the description sentence with one person is affirmed at first. If the character is not the extracted object, omit the sentence. If the character is the extracted object, attributes are extracted and put into different attribute lists in terms of marks. For example:

Example Sentence 17. 1943 年 11 月/t ，/w 白志东/nr 出生/bir 于/p 河北省/sh 乐亭县/sx。/w

Translation: In November, 1943, Bai Zhidong was born in Leting County, Hebei Province.

According to the feature word “出生(birth)” and attribute marks, the attributes of “1943 年 11 月(Nov. 1943)”, “河北省(Hebei province)” and “乐亭县(Laoting county)” are put into such attribute lists as date of birth, province of birth and city of birth(including towns and villages) of the extracted person “白志东”.

4.2.2 The description sentences with more persons

Attribute ownership decision in the description sentences with more persons turns out be the challenge of this evaluation task. For example:

Example Sentence 18. 李济深升为军长，陈济棠升任第十一师师长。（陈济棠）

Translation: Li Jishen was promoted to an army corps commander, and Chen Jitang was promoted to be the commander of eleventh division. (Chen Jitang)

In this sentence, “军长(army commander)” is the title of “李济深”, while “师长(divisional commander)” is the title of “陈济棠”, a person to be extracted. Attribute ownership decision requires us to correctly recognize “陈济棠” and then extract it. We mainly employ minimum slicing with the co-occurrence of the extracted person and the attribute and the nearest distance principle to decide attribute ownership, which will be expounded below.

(1) minimum co-occurrence slicing

When the person and the attribute co-occur in the same grammatical unit as minimum as possible, and there is only one person, the attribute belongs to the person. For example:

Example Sentence 19. 1947 年 4 月冀察热辽军区部队改编为东北民主联军第八纵队，黄永胜任司令，丁盛任二十四师师长，之后参加了辽沈战役。

Translation: In April, 1947, the troops of Ji-Cha-Re-Liao military region were reorganized as the 8th Army of the Northeast Democratic Coalition Force. Huang Yongsheng became the commander, and Ding Sheng took the post of commander of 24th division, then they took part in the Liaoning-Shenyang Campaign.

Example Sentence 20. 1935 年，蒋中正调张学良东北军剿共，西安出现以西北剿匪总司令部副总司令张学良、西安绥靖公署主任杨虎城和陕西省政府主席邵力子为首三种势力并存局面。（杨虎城）

Translation: In 1935, Chiang Kai-shek dispatched Zhang Xueliang's Northeast Army to conquer the communist power. There coexisted three powers in Xi'an, ie the power of Zhan Xueliang, who was the Vice Commander in chief of Northeast Anti-communist Army; the power of Yang Hucheng, who was the director of Xi'an Appeasement Administrative Office; and the power of Shao Lizi, who as the governor of Shaanxi provincial government. (Yang Hucheng)

In the two clauses of Example Sentence 19, “黄永胜任司令” and “丁盛任二十四师师长” means the title of “司令(commander)” belongs to “黄永胜”, while the title of “师长(divisional commander)” belongs to “丁盛”. In Example Sentence 20, “副总司令张学良”, “主任杨虎城” and “主席邵力子” show that the person and the attribute co-occur in the same subject-predicate phrase.

(2) the nearest distance principle

When there is a long distance between the person and the attribute, and at the same time, there are more persons in the sentence, the attribute belongs to the person with the nearest distance.

Example Sentence 21. 钱三强的父亲钱玄同是中国近代著名的语言文字学家。（钱三强）

Translation: Qian Sanqiang's father, Qian Xuantong is a famous modern Chinese linguist.

(Sanqiang Qian)

Example Sentence 22. 我从小就知道江泽涵是北京大学一位鼎鼎大名的数学教授，却无缘见面，但他们的堂姐江冬秀我却在孩童时就见过。（江泽涵）

Translation: When I was young, I got to know that Jiang Zehan is a famous math professor of Peking University, but I had no luck to meet him; but I'd seen their cousin Jiang Dongxiu during my childhood. (Zehan Jiang)

Example Sentence 23. 薛万彻的二哥薛万淑，也战功显赫，历任右领军将军、梁郡公、畅武道行军总管。（薛万彻）

Translation: Xue Wanche's second brother, Xue Wanshu also made daring military exploits, who used to be a general of the right wing, Duke of Liang Jun, and Commander in Chief of Changwudao.(Wanren Xue)

In Example Sentence 21, the title “语言文字学家(linguist)” belongs to “钱玄同” instead of “钱三强”, for the distance between the title “语言文字学家(linguist)” and the person “钱玄同” is smaller. The situations in Example Sentence 22 and Example Sentence 23 are also like this. It should be noted that the nearest distance principle is not always effective, as in the following example sentence.

Example Sentence 24. 中共四大后，彭述之以中央委员身份接替多病的蔡和森担任中央宣传部长，为了工作方便，蔡和森夫妇、彭述之夫妇等人一起住在宣传部的寓所。

Translation: After the 4th National Congress of CPC, as a member of the Central Committee of CPC, Peng Shuzhi take the place of Cai Hesen, who was sick, to be the minister of the State of Central Propaganda Ministry. In order to facilitate the work, both Hesens and Shuzhis lived in the apartments of Propaganda Ministry.

In Example Sentence 24, the title “部长(minister)” belongs to “彭述之”, the person which has a longer distance. This sentence needs deeper syntax or semantic analysis, which is a little difficult to process at present.

4.2.3 anaphora resolution of person pronouns[2]

As for anaphora resolution in the description sentences with more persons, we mainly refer to the methods in (Wang, 2001; Wang, 2005). The extracted person is known, so its designation and sex can be annotated in advance, which facilitates anaphora resolution. For example:

Example Sentence 25 & 26：1940 年，钱三强取得了法国国家博士学位，又继续跟随第二代居里夫妇当助手。1946 年，他与同一学科的才女何泽慧结婚。

Translation: In 1940, Qian Sanqiang obtained his French national doctorate, and then he continued to follow Curies, the junior, as an assistant. In 1946, he married the talented girl He Zehui, who was learning the same subject.

As in Example Sentence 25, “居里夫妇” is plural, “他(He)” in Example Sentence 26 refers to “钱三强” in the preceding sentence, which is a male name in singular form.

4.3 Attribution extraction flowchart

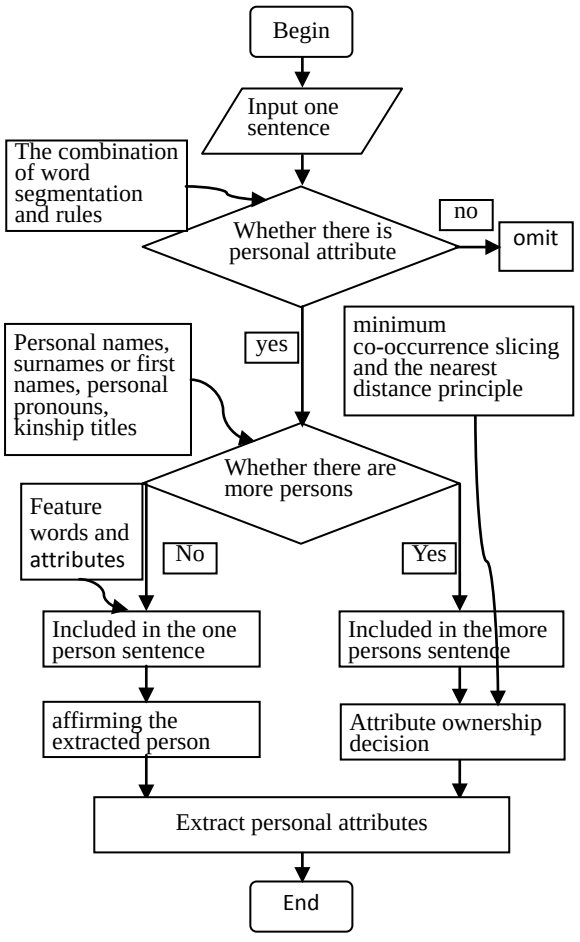


Fig. 5 the Flowchart of Personal Attribution Extraction

5 Experimental results

In this bakeoff, the performance of 6 groups attending the competition are shown in Table 1.

[2]Since there are few cases of reverse anaphora, it has not been considered in this text.

Our system is named as CASIA_CUC_PAES.

Table 1. The lenient and strict evaluation results

Team Id	lenient SF_Value	strict SF_Value
CIST-BUPT	0.363235496	0.352206490
ICTNET_002	0.277775207	0.273884523
WZ_v4	0.004311033	0.002491385
BLCU-yudong	0.308706661	0.292608955
Result-BUPT	0.071467108	0.035979785
CASIA_CUC_PAES	0.507388780	0.489505010

According to the evaluation results, our system achieves 0.507388780 and 0.489505010 respectively in the lenient evaluation results and the strict evaluation results of SF_Value in CIPS-SIGHAN2014 Bakeoff, which turns out to be the best. The fact has shown that our system is effective. However, 50 percent of SF_Value implies that there is still room to increase the system's efficiencies. The system performance could be improved in 3 aspects:

1. to establish the word segmentation system specific for personal attribute extraction.
2. to establish grammatical knowledge system regarding personal attribute extraction, For example, "我父亲住在北京(My father lived in Beijing)" is different from "我和父亲住在北京(My father and I live in Beijing)", with "我父亲" as a modifier-head construction in the former and "我和父亲" as a parallel construction in the latter.
3. to establish semantic knowledge system regarding personal attribute extraction, For example, in the sentence of "凯利与女演员劳里·莫顿结婚后居住于 Goatstown.(After wedding, Kerry and actress, Laurie Morton settled in Goatstown.)", certain semantic knowledge is needed to correctly extract the information that Laurie Morton is Kelly's wife.

6 Conclusion

This bakeoff is full of challenges with a number of personal attributes to be extracted. CUCBst, the word segmentation software, plays a significant role in named entity recognition, which provides a solid foundation for attribute extraction. The strategy of sentence classifications is employed in attribute ownership decision, which, though cannot solve all the problems, simplifies analyses. This strategy plays a role in improving precision in attribute ownership decision.

References

Bikel, D., Castelli, V., Florian, R., & Han, D. J. 2009, November. Entity linking and slot filling through statistical processing and inference rules. In *Proc. TAC 2009 Workshop*.

Burman, A., Jayapal, A., Kannan, S., Kavilikatta, M., Alhelbawy, A., Derczynski, L., & Gaizauskas, R. 2012. USFD at KBP 2011: Entity linking, slot filling and temporal bounding. *arXiv preprint arXiv*:1203.5073.

Chen Ping. 1987. Discourse Analysis of Chinese Zero Anaphora. *Studies of The Chinese Language*, 5: 363-378.

Kong Fang, Zhou Guodong, & Zhu Qiaoming. 2010. Survey on Coreference Resolution. *Computer Engineering*, 36(8): 33-36.

Liao Xiantao, Yu Haibin, & Qin Bing, Liu Ting. 2004. HMM combined with automatic rules-extracting for Chinese Named Entity recognition. *The second national student Workshop on Computational Linguistics*. Beijing:[s. n.], 2004: 232-237.

McNamee, P., & Dang, H. T. 2009, November. Overview of the TAC 2009 knowledge base population track. In *Text Analysis Conference (TAC)* (Vol. 17, pp. 111-113).

McNamee, P., Dang, H. T., Simpson, H., Schone, P., & Strassel, S. 2010, May. An Evaluation of Technologies for Knowledge Base Population. In *LREC*.

Sun Maosong, & Gao Haiyan. 1995. Identifying Chinese Names in Unrestricted Texts. *Journal of Chinese Information Processing*, 9(2), 16-27.

Sun, A., Grishman, R., Xu, W., & Min, B. 2011. New York University 2011 system for KBP slot filling. In *Proceedings of the Text Analytics Conference*.

Wang Houfeng, & Tingting He. 2001. Research on Chinese Personal Pronoun Anaphora Resolution. *Chinese Journal of Computers*, 24(2), 136-143.

Wang Houfeng, & Zheng Mei. 2005. Robust Pronominal Resolution within Chinese Text. *Journal of Software*, 16(5), 700-707.

Wang Houfeng. 2002. Computational Models and Technologies in Anaphora Resolution. *Journal of Chinese Information Processing*, 16(6), 9-17.

Wang Houfeng. 2005. On Anaphora Resolution

within Chinese Text. *Applied Linguistics*, (4), 113-119.

Xu Jiujiu. 2003. Anaphora in Chinese Texts. *China Social Sciences Publishing House*.

Ye Zheng, Lin Hongfei, & Shu Sui. 2007. Person Attribute Extracting Based on SVM. *Journal of Computer Research and Development*, (z2): 271-275.

Yu Hongkui, Zhang Huaping, & Liu Qun. 2006. Chinese named entity identification using cascaded hidden Markov model. *Journal on Communications*, 27(2).

Zhao Jun, Huang Changning. 1999. A Transformation-Based Model for Chinese BaseNP Recognition. *Journal of Chinese Information Processing*, 13(2): 1-7.

Zhuang Shouqiang. 1997. Difference and Relation Between Features and Attribute and its Significance in Scientific Research. *Studies In Dialectics of Nature*, 13(11): 44-48.