Personal Attributes Extraction Based on the Combination of Trigger Words, Dictionary and Rules

Kailun Zhang, Mingyin Wang, Xiaoyue Cong, Fang Huang, Hongfa Xue, Lei Li,Zhiqiao Gao

Beijing University of Posts and Telecommunications,
China，100876
kailun0315@qq.com, wmy512@qq.com, cxy0105@bupt.edu.cn

xprince.hf@gmail.com, xuehongfa@vip.qq.com, leili@bupt.edu.cn, 526804113@qq.com

Abstract

Personal Attributes Extraction in Unstructured Chinese Text Task is a subtask of The 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2014). In this report, we propose a method based on the combination of trigger words, dictionary and rules to realize the personal attributes extraction. We introduce the extraction process and show the result of this bakeoff, which can show that our method is feasible and has achieved good effect.

Keywords: Unstructured Chinese Text, Personal Attributes Extraction, Trigger Words, Dictionary, Rules

1 Introduction

In recent years, with the development of Internet, masses of information provide the majority of Internet users with a lot of convenience. However, with the increase of amount of information, screening redundant information and seeking for the knowledge which users really want from a lot of unstructured texts is getting more and more difficult. For example, when we search for the details of someone, general search engines usually return a number of pages, and we must identify these pages one by one even if we just need a little of them. Therefore, extracting personal attributes from unstructured texts has become a very important task. Personal attributes extraction in unstructured Chinese text task is designed to extract person specific attributes, such as date of birth, spouse, husband, children, education, or title etc. from unstructured Chinese texts. The corresponding techniques play an important role in information extraction, event tracking, entity disambiguation and other related research areas.

In our report, a method based on the combination of trigger words, dictionary and rules to realize the personal attributes extraction is introduced. We build a basic framework including trigger words, dictionaries and rules that relative to the task to extract personal specific attributes. In Section 2, we introduce two basic methods about information extraction and several recent researches on this theme while the detailed description of the task is represented in Section 3. In Section 4, we give the step to build our basic model of extraction. We talk the main framework in Section 4.1. Then from Section 4.2 to Section 4.4, we describe the process to build trigger word table, attribute dictionaries and personal attribute rules one by one in a detailed way. We show the evaluation metrics and the final experiment results in Section 5 to prove the feasibility of our method. In Section 6, we point out the shortage of our system and propose some suggestions to improve our model and then make a conclusion.

2 Related Works

Rule-based methods and statistics-based ones are two main ways of information extraction at present. Information extraction based on the rules is a two phase process consists of learning and applying, including the study of rules and the application of using rules for target information extraction. Information extraction rules mainly come from the target context in constraint environment. As long as finding the constraint information which can meet the rules in the text, we could also find the target extraction information. Thus, learning and extracting the rules themselves becomes the key point to the rule-based information extraction. As for the method of statistics-based, its accuracy is generally low, but it has good portability to this extraction problem. Some statistics models have strong statistical theory basis and wholesome training algorithms such as HMM and CRFs and so on. However, statistics-based information extraction requires a large amount of labeled training data.

Currently, there are not many references about the personal attributes extraction and there is no more mature system to solve this problem. However, personal attributes extraction has a very close relation to the information extraction, and personal entities also belong to the category of the entity. So, to a certain extent, the entity relation extraction method can also be applied to personal attribute extraction. Ye [1] and some other researchers treated the personal attribute extraction as a specific application in the entity relationship extraction. They use the “Hownet” to acquire the trigger words which can describe the personal attributes, then change the relationship between trigger words and names into a classification problem. Their solution needs manual

Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 114–119,
Wuhan, China, 20-21 October 2014

labeled data during classifier training and is under the help of semantic resource. Wang [2] and some other researchers put forward a relationship judgment algorithm which is based on the semantic similarity between the current tuples and the relationship set to filter and classify the relational tuples that are extracted according to the pattern, using Wikipedia as a knowledge database. This is under the foundation of extraction model of sentence groups such as blocks and named entity recognition marker. Wang [3] and others tried to use the method of knowledge engineering to extract personal attributes. They sum up some rules manually under the foundation of mass analysis about web texts and researches in natural language processing and then built a pattern repository to do the match. Yu [4] adopted the way of using trigger words and classifier to exact personal basic information, and carried out a character search engine based on the stored exaction information.

3 Task Descriptions

In this task, there are 25 predefined personal attributes to be extracted, including alternate_names, date_of_birth, age, country_of_birth, stateorprovince_of_birth, city_of_birth, date_of_death, country_of_death, stateorprovince_of_death, city_of_death, coutriea_of_residence, stateorprovince_of_residence, cities_of_residence, title, member_of, employee_of, religion, spouse, children, parents, siblings, other_family, charges, cause_of_death and schools_attended. The testing data are provided by a series of folders which are named after people whose attributes need to be extracted. In each folder, a XML document of Wikipedia and some unstructured Chinese texts about the person are included. Except for the actual attribute values, the extraction results should also contain the source documents that the values come from and their positions in the documents. For the attributes that are already located in the tags of “Facts” in the document of Wikipedia, we do not need to extract them repeatedly. For those attributes whose values are not unique, such as parents, children and the residence of cities, it is responsible for us to extract all probable attribute values.

4 Methods

Before the selection of methods to extract, we’ve analyzed the attributes to be extracted, the sample data and also the testing data provided by the conference carefully. Because we don't have enough data as the training data, and it requires quantities of work to collect and label the training data artificially, we gave up the extraction method based on statistics. While, through the observation of a large number of Wikipedia pages and personal information, we found that most of the attributes have a great similarity in the expression and discipline. Therefore, what we use is a method that combines the trigger words, dictionaries and rules together to achieve the task of personal attributes extraction.

4.1 Basic Framework

As shown in Figure 1, the architecture includes several parts:

1. The test corpus is provided by the conference. The corpus includes several XML files about persons whose personal attributes are to be extracted, containing the persons’ Wikipedia records, and a number of unstructured documents relating to the persons.

2. Build attributes trigger words. The trigger words are aimed to narrow down the extraction scope, such as birth date and place of birth appears in sentences containing “出生”(birth) or “生于”(born).

3. Build attributes dictionary. The dictionary is in the view of the state, province, and school, the cause of death and some similar fixed attributes or some attributes which could be extracted by dictionary lookup directly.

4. Build attributes extraction rules. We sum up the general characteristics of the attributes from the corpus using the combination features of word segmentation, part-of-speech (POS) tagging, named entity recognition (NER) and sentence parsing. Then we formulate the rules of grammar corresponding to these characteristics respectively. As a result, we can use these rules in the process of personal attributes extraction respectively.

5. Extract the attributes information. Extract attributes from the input unstructured documents according to the rules and structure of the dictionary.

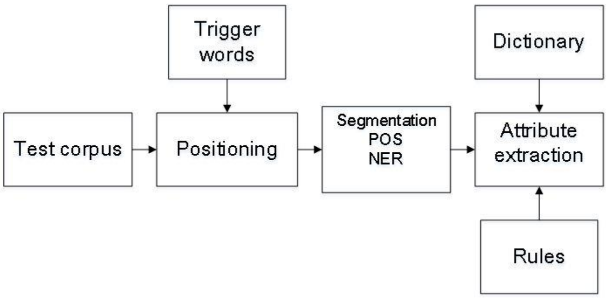

Figure 1 System Framework

4.2 Build Trigger word Table

So-called trigger word refers to a particular attribute extraction having the effect of location and identification that can activate the extraction task. When a sentence contains trigger words in a certain document, it could trigger the corresponding attribute extraction task in the sentence, so that the scope of the attribute extraction would be greatly narrowed. In this work, by analyzing the text characteristic and the description of the Chinese language style, we built trigger word sets for part of the corresponding attributes, while the attributes without trigger words require full range extraction in document. Trigger word table is shown in Table 1.

Table 1 Trigger word Table of Personal Attributes Extraction

Name of Attributes	Trigger word Set
alternate_names	本名（autonym），原名(primitive name），曾用名(used name)，中文名(Chinese name)，英文名(English name)，日文名(Japanese name)，全名(full name)，谥(posthumous title), 号称(known), 字(styled)，尊名(name being) , etc.
date_of_birth, country_of_birth, city_of_birth, stateorprovince_of_birth	出生 (birth), 生于(born)
age	岁(age)
date_of_death, country_of_death, city_of_death, stateorprovince_of_death, cause_of_death	逝世 (die), 去世(pass away)，死于 (die of), 卒于(die in)，殉命于(to perish)，病死 (die in one’s bed), 病故 (die of illness) , and the year and date extracted from the record in the <date_of_death> tag.
schools_attended, countries_of_residence, citis_of_residence, statesorprovince_of_residence,	就读 (attend), 受教育(educated by)，选修(elective course)，学习(study), 毕业(graduate), 转读(transfer)，读书(read), 硕士(master)，博士 (PhD), 学士(bachelor)，本科(undergraduate), 迁居(move), 流亡 (exile), 移居(migrate)，定居(settle), 故居(hometown), 长大(grow up), 多年 (many years), 几年(several years), 居住(live)，任(appoint)，创作出(create)，从事(be occupied in)，工作(work)
title	担任(take charge of), 历任(successively held the posts of), 成为 (become), 任(appoint)，为(as), 当(work as), 封(confer) , etc.
member_of, employee_of	进入 (enter into), 签约(sign a contrast), 打工 (work part-time)，任教(work as a teacher)，旗下(subordinate), 受聘(offered appointment)，晋升 (promote), 任命(nominate), 升(promote)，聘(employ)
religion	信奉 (believe in), 信仰(belief), 信 (believe)，徒(follower)
spouse	配偶(spouse)，妻(wife)，结婚 (marriage)，丈夫(husband)，完婚 (get married)，太太 (Mrs.), 夫人(madam)，遗孀(widow)，嫁，娶(take to wife), 结为伉俪(married couple)，奉子成婚(shotgun marriage), 王后(Queen), 皇后(King), etc.
parents	父亲(father), 母亲(mother), 其父(one’s father), 其母(one’s mother)，庶母(concubine of one's father), 妈妈(Mama)，随父(following one’s father)，随母(following one’s mother), etc.
children	儿子(son), 女儿(daughter), 子女(children)，之子(one’s son)，之女(one’s daughter)，幼女(infantile daughter)，幼子(infantile son)，长子(eldest son), 长女(eldest daughter), 次子(second son), 次女(second daughter), 二子(second son), 三子(third son), 四子(fourth son), etc.
siblings	哥哥(older brother)，弟弟(younger brother), 姐姐(older sister)，妹妹(younger sister), 长兄(eldest brother), 姊姊(sister)，大姊 (eldest sister), 小妹(youngest sister), 二哥(second elder brother)，兄弟(brother), etc.
other_family	祖父 (grandfather), 祖母(grandmother)，叔叔(uncle), 表兄(elder male cousin), 表姐(elder female cousin)，妹夫(brother-in-law)，同族兄弟(Cousins), 岳父(father-in-law), 侄 (nephew), 甥 (nephew), 舅 (mother’s brother)，堂姐(elder female cousin), 堂兄 (elder male cousin), 内兄(brother-in-law) , etc.
charges	Words containing “罪”(crime)

4.3 Build Attribute Dictionary

We built attribute dictionary aiming at national, provincial or state, city, school, etc. for those attributes, which can be extracted directly by dictionary lookup. Compared to the rules, dictionary extraction is more convenient and with higher accuracy. For part of attributes, we built 8 dictionaries referring to the country, school, religion etc., as shown in Table 2.

Table 2 Dictionary of Personal Attributes Extraction

Name of Attributes	Content of Dictionary
country_of_birth, country_of_death, countries_of_residence	The full names and abbreviations of all the countries
city_of_birth, city_of_death, citis_of_residence	The cities of all countries and the towns or areas of China
stateorprovince_of_birth, stateorprovince_of_death, statesorprovince_of_residence	The states or provinces of all countries
schools_attended	All schools and colleges throughout the world
religion	All religions
cause_of_death	Common cause of death, such as “自杀”(suicide), “枪决”(execute by shooting), etc
charges	Common crime, such as drug trafficking, debt, etc.
title	Words about job, rank, field position and ancient official position, and the title attribute from the sample data

4.4 Build Personal Attribute Rules

Rules are very important for the proposed personal attributes information extraction. Its quality directly decides the effect of information extraction. While we were studying the personal attributes, we found that the expression of same attributes have a lot of similarities. Based on the similarity, in combination with word segmentation, part-of-speech tagging, and named entity recognition, we built rules for each corresponding attribute. Rule sets are shown in Table 3.

Table 3 Rules of Personal Attributes Extraction

Name of Attributes	Rules
alternate_names	The words after the trigger words connected with punctuation marks; The recent word tagged by “NN” after the trigger words; The quoted words after the trigger words
date_of_birth	Generated in advance all the regular time format templates, and match the time format in the first sentence containing trigger words as the result
country_of_birth, city_of_birth, stateorprovince_of_birth	Match the corresponding dictionary in the first sentence containing the trigger words
age	extract numbers followed by the "岁", taking the maximum as a result; Add specific rules to extract, For the Chinese digital age, such as “六十岁”
date_of_death	Match time format in the sentence containing the trigger words as a result when the content of <date_of_death> tag is empty.
country_of_death, city_of_death, stateorprovince_of_death	Match the corresponding dictionary in the sentence containing the trigger words
cause_of_death	Match the corresponding dictionary in the sentences containing trigger words; Search for the string with a tag sequence of NN or NN + NN + VV or NN + NN or NN + VV or NN + VA after the “由于” or “因” whose tag is “P” with a distance less than five words until meeting punctuation.
schools_attended, countries_of_residence, citis_of_residence, statesorprovince_of_residence,	Match the corresponding dictionary in the sentence containing the trigger words
title	Match the title dictionary backward in the phrase containing trigger words or the character name; The recent word tagged by “NN” after the phrase with the structure of the trigger words or character name + "是" ; match the title dictionary in all the sentences containing the character name when the query failed.
member_of, employee_of	The chunks tagged by “ORG” after named entity recognition in the sentences containing the

	trigger words or title attribute; Search for the recent chunk tagged by “NP” in phrase containing trigger words, bidirectionally; Mark the results containg " 会 ", " 军 ", " 队 " as member_of atrribute value, the rest as employee_of attribute values
religion	Match the religion dictionary in the sentences containing trigger words
spouse, parents, children, siblings	The chunks tagged by “PER” after named entity recognition in the sentences containing the trigger words, rejecting the character name
other_family	The chunks tagged by “PER” after named entity recognition in the sentences containing the trigger words, rejecting the character name or the name marked by other attributes.
charges	match the corresponding dictionary in sentences containing the character name; Search for the string with a tag sequence of VV or AD+VV before the trigger word. The string between the phrase and the trigger word is the value.

5 Experiments

This work is designed to extract person specific attributes from unstructured Chinese texts. The testing date contains 323 documents about 90 persons, including 233 documents to extract attributes and 90 documents from Wikipedia records. The organizer of Personal Attributes Extraction in Unstructured Chinese Text Subtask of The 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2014) takes the same evaluation metrics adopted in the slot filling of TAC KBP. Deails of the result is presented in [5].

Single attributes evaluation metric

$$Score_{single} = \frac{NumCorrect}{NumSingleSlot}$$

When NumCorrect is zero, we set NumCorrect to 1.0;

List attributes evaluation metric

$$ListSlotValue = \frac{({F_\beta}^2 + 1) * IP * IR}{{F_\beta}^2 * (IP + IR)}$$

$$F_\beta = 2 \ to \ weight \ precision \ over \ recall$$

$$IP = Instance \ precision$$

$$IR = Instance \ recall$$

$$Score_{list} = \frac{\sum ListSlotValue}{NumListSlots}$$

When both IP and IR are zero, we set ListSlotValue to 0.0;

Overall evaluation metric

$$SF_{value} = \frac{1}{2}\left(Score_{single} + \ Score_{list}\right)$$

We use the average of single attributes evaluation score and list attributes evaluation score as the final evaluation score. In the evaluation, both the lenient evaluation and strict evaluation are performed. In the strict evaluation, all instance attributes are compared to the answers while in the lenient evaluation, the offsets of the string from the beginning word to the ending word are ignored. Table 4 and Table 5 give the results for lenient evaluation and strict evaluation, respectively. Note that there are 6 teams participated in this bakeoff, as shown in the first column of Table 4 and Table 5, in which our team is called CIST-BUPT.

Table 4 the Lenient Evaluation Results

Team	Single Score	List Score	SF_Value
CIST-BUPT(Ours)	0.562770563	0.163700429	0.363235496
ICTNET_002	0.350649351	0.204901063	0.277775207
WZ_v4	0.004329004	0.004293061	0.004311033
BLCU-yudong	0.428571429	0.188841894	0.308706661
Result-BUPT	0.121212121	0.021722095	0.071467108
CASIA_CUC_PAES	0.670995670	0.343781890	0.507388780

Table 5 the Strict Evaluation Results

Team	Single Score	List Score	SF_Value
CIST-BUPT(Ours)	0.549783550	0.154629430	0.352206490
ICTNET_002	0.350649351	0.197119695	0.273884523
WZ_v4	0.004329004	0.000653766	0.002491385
BLCU-yudong	0.411255411	0.173962498	0.292608955
Result-BUPT	0.060606061	0.01135351	0.035979785
CASIA_CUC_PAES	0.645021650	0.33398837	0.489505010

We can see that our method has achieved good results, ranking the second place in the six teams. The results fully show that the method based on the combination of trigger words, dictionary and rules is feasible to some extent, and the trigger words and rules we formulated have performed well.

But there are still some problems in our method. The list attributes evaluation score is far lower than the single attributes evaluation score, which shows that we possibly have missed a lot of instances. And when considering the offsets of the extracted string, both the single attribute and list attributes evaluation score declined. This indicates that there are some errors, for example, the attribute value is correct but the source or object is wrong. In future work, we need to develop special improved strategies to extract more accurate results.

6 Conclusions and Future Work

In this report, we proposed a method based on the combination of trigger words, dictionary and rules to extract person specific attributes from unstructured Chinese texts. The trigger words can narrow the scope of extraction and then they are combined with specific dictionary lookup and extraction rules to implement the extraction of 25 person specific attributes.

Given the limited time and the first try in this kind of bakeoff, our system still has some shortages to be improved. For example, in the case of “Missing Words”, we can specify the rules or collect and tag data artificially in order to get more training data and then use the method of machine learning to extract person attributes. On the other hand, to improve the case of “Incorrect Words”, we plan to increase the judgment of the subject in one sentence so that we can avoid the situation that the attributes we extract belong to other people. Otherwise, we can also try to make more specific rules for the place names which occurs in schools or organizations to reduce their effects to those related attributes about place.

We believe that if we do some improvements to our system as above, we can get a more accurate extraction result. And we are also looking forward to developing more formal and more relatively complete machine learning algorithms and rules to realize the extraction of person specific attributes in unstructured Chinese with less human labor in the loop.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 61202247, 71231002, 61202248 and 61472046; EU FP7 IRSES MobileCloud Project (Grant No. 612212); the 111 Project of China under Grant B08004; Engineering Research Center of Information Networks, Ministry of Education; the Fundamental Research Funds for the Central Universities under Grant2013RC0304; the project of Beijing Institute of Science and Technology Information.

References

[1] Zheng Ye, Hongfei Lin, Sui Su, Jingjing Liu, Person Attribute Extracting Based on SVM, Journal of Computer Research and Development[J], 2007, 44:271-275

[2] Quanjian Wang, Fang Wang. Wikipedia Based Name and Resume Information Extraction[J]. Computer Applications and Software，2011, 27(7):170-174.

[3] Ying Wang. Research on Web Information Extraction Applied to Chinese Name Search Engine. Lanzhou University. Thesis, 2006.

[4] Manquan Yu. Research on Knowledge Mining in Person Tracking. Institute of Computing Technology, Chinese Academy of Science. Dissertation, 2006.

[5] CLP 2014 Shared Task: Personal Attributes Extraction in Chinese Text. Evaluation Report for The 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2014), 2014.