Personal Attributes Extraction in Chinese Text Bakeoff in CLP 2014: Overview

Ruifeng Xu, Shuai Wang, Feng Shi
Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School, Harbin Institute of Technology, China
xuruifeng@hitsz.edu.cn

Jian Xu
Department of Computing, The Hong Kong Polytechnic University, Hong Kong
csjxu@comp.polyu.edu.hk

Abstract

This paper presents the overview of Personal Attributes Extraction in Chinese Text Bakeoff in CLP 2014. Personal attribute extraction plays an important role in information extraction, event tracking, entity disambiguation and other related research areas. This task is designed to evaluate the techniques for extracting person specific attributes from unstructured Chinese texts, which is similar to slot filling, but focuses on person attributes. This task brings some challenges issues because Chinese language contains some common words and lacks of capital clues as in English. The task organizer manually constructs the query names and corresponding documents. The value/presence of the texts corresponding 25 pre-defined attributes are annotated to construct the training and testing dataset. The bakeoff results achieved by the participators show the good progress in this field.

1 Introduction

Personal Attributes Extraction in Chinese Text Task is designed to evaluate the techniques for extracting person specific attributes, such as birth date, spouse, children, education, and title etc. from unstructured Chinese texts. These techniques play an important role in information extraction, event tracking, entity disambiguation and other related research areas.

Slot filling task has been proposed as one of shared tasks in the TAC KBP workshop since 2009 [1]. Generally speaking, the mainstream techniques for slot filling and person attributes extraction may be camped into two major approaches, namely: Rule-based approach and statistics-based ones [2,3,4]. Rule-based approach normally defines the extraction rules manually or learns the rules automatically. The rules play the key role in this approach. As long as finding the constraint information which matches the rules in the text, the system may extract the target extraction information. As for the statistics-based approach, it has good portability to this extraction problem. Several statistics machine learning models such as Hidden Markov Model (HMM) and Condition Random Fields (CRFs) are employed. The shortcoming for this approach is that it requires large amount of training data which is always unavailable.

Currently, there are limited existing works on personal attributes extraction in Chinese text. Comparing to the works on English, the characteristics of Chinese language including the Chinese word segmentation, the confusion of named entity with common words, lack of capital clues bring more difficulties for person attributes extraction in Chinese.

The task of person attributes extraction in Chinese text in CLP 2014 bakeoff is designed on the basis of the slot filling task in the TAC KBP workshop [1]. The task organizer provides a collection of documents corresponding to a target person and a knowledge base which contains partial list of attributes for the person. Participants are required to extract additional attributes from the collections of documents. The task is similar to the slot filling, but it focuses on person attributes extraction. Furthermore, the collection of documents is not limited to the news corpus.

Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 108–113,
Wuhan, China, 20-21 October 2014

2 Task Definition

2.1 Task description

The Personal Attributes Extraction in Chinese Text Task is motivated by a component of a full slot filling (SF) system. This task focuses on the refinement of output from Chinese slot filling systems. Especially, personal attributes extracted from the unstructured text is useful for the construction of Chinese knowledge graph.

In this task, the participants are provided a set of document collection in several person name folders. In each folder, source documents named as XXX_Ti.xml and Wikipedia knowledge base named as XXX 维基百科记录.xml are given. The Wikipedia knowledge base for each person is an XML document, in which attributes are located in the tags of Facts. In addition, unstructured text for that person is also provided with the wiki_text tag. Example 1 gives a sample record in Wikipedia knowledge base.

<entity wiki_title="周强" type="PER" id="" name="周强">
<facts class="Infobox">
<fact name="nationality">中国</fact>
<fact name="birthdate">1960年4月</fact>
<fact name="education">西南政法大学</fact>
</facts>
<wiki_text>周强（1960年4月－），湖北黄梅人，西南政法大学民法专业毕业，法学硕士。
</wiki_text>
</entity>

Example 1: A Sample Wikipedia knowledge base.

The extraction task focused on extracting values for a set of pre-defined attributes (“slots”) for target person entity from given source documents. Given an entity, the system is required to extract the correct value(s) for that pre-defined attribute from source documents and return the slot filler together with its provenance, which is a set of text spans from source document that justify the correctness of the slot filler. The extraction system need not extract the attribute values given in the Wikipedia knowledge base.

2.2 Dataset preparation

The person names are manually selected from the web, in which 10 person names are used in training dataset and 90 person names, including 48 names for Chinese person and 42 names for foreign person are used in testing dataset. The corresponding knowledge base is constructed from Wikipedia person entity while the source documents in each folder are constructed based on search engine output with manually selection.

The personal attributes are categorized as being Person (PER) slots based on the type of entities about which they seek to extract information. The attributes are also categorized by the content and quantity of their fillers [5].

2.2.1 Attribute slot content

Attribute slot content are divided into three categorizations, namely Name, Value, or String.

Name slots are required to be filled by the name of a person. Name slots including the alternative name, spouse name, city of birth, country of death and so on. The detailed slot descriptions are given in the Personal Attributes Extraction in Chinese Text Task website.

Value slots are required to be filled by either a numerical value or a date such as age and birth date. The numbers and dates in these fillers can be spelled out (forty-two; December 7, 1941) or written as numbers (42; 12/7/1941).

String slots are basically a “catch all”, meaning that their fillers cannot be neatly classified as names or values. The text excerpts (or “strings”) that make up these fillers can sometimes be just a name, but are often expected to be more than a name. The typical string slots including cause of death and religion.

2.2.2 Attribute slot quantity

Slots are labeled as Single-value or List-value based on the number of fillers they can take. Since one slot may have different representations, participant is required to extract all of these representations.

Single-value slots can have only single filler. While most single-value slots are obvious (e.g., a person can only have one date of birth), some may be less apparent.

List-value slots can take multiple fillers as they are likely to have more than one correct answer in the source data. For example, people may have multiple children, employers, or alternate names.

2.2.3 Attribute Table

The following table of all 25 pre-defined attribute slots and their categorizations is given below.

Slot name	Content	Quantity
Alternate Names	Name	List
Children	Name	List
Cities of Residence	Name	List
City of Birth	Name	Single
City of Death	Name	Single
Countries of Residence	Name	List
Country of Birth	Name	Single
Country of Death	Name	Single
Other Family	Name	List
Parents	Name	List
Schools Attended	Name	List
Siblings	Name	List
Spouses	Name	List
Stateorprovince of Birth	Name	Single
Stateorprovince of Death	Name	Single
Statesorprovinces of Residence	Name	List
Age	Value	Single
Date of Birth	Value	Single
Date of Death	Value	Single
Cause of Death	String	Single
Charges	String	List
Religion	String	List
Title	Name	List
Member of	Name	List
Employee of	Name	List

Table 1. Attribute slots

In this task, the organizer collects the source documents under each person name by using the search engine. Using the person name and the related attribute names as the query to search on the Internet, the top N high quality web pages are manually selected as the source documents. During the set construction, the organizer avoids to the attribute slots overlapping between different source documents. Table 2 gives the statistical information for source document.

Sets	Max	Min	Average	Total
Train set	4	1	2	24
Test set	5	1	2	235

Table 2. Statistical information of source documents

The instance means one person's attribute slot appears in one source document. Table 3 lists the detail information about the instance number of one related person attribute in one source document.

attributes	Max	Min	Average
Single	6	0	1
List	47	0	1

Table 3. Instances in source documents

As mentioned above, the person attributes are divided into two categorizations: Single and List. The total instance numbers for the two categorizations in the training set and testing set are shown as follows.

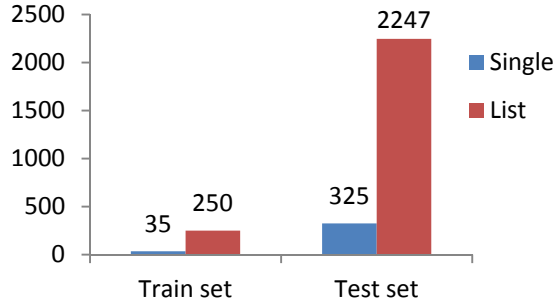

Figure 1. The instance numbers in the training set and testing set.

3 Evaluation Metrics

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 offset string_begin and string_end are ignored. The detail evaluation metrics are shown as follows.

3.1 Single Attributes Evaluation Metric

$$\text{Score}_{\text{single}} = \frac{\text{NumCorrect}}{\text{NumSingleSlot}} \quad (1)$$

When numCorrect is zero, the numCorrect is set to 1.0;

3.2 List Attributes Evaluation Metric

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

$$Score_{list} = \frac{\sum ListSlotValue}{NumListSlots} \quad (3)$$

When IP is the instance precision and IR is the instance recall, in the evaluation we set the weight $F_\beta = 2$, and when both IP and IR are zero, we set the ListSlotValue to zero;

3.3 Overall Evaluation Metric

$$SF_{value} = \frac{1}{2}\left(Score_{single} + Score_{list}\right) \quad (4)$$

The overall evaluation metric is the average of single attributes evaluation score and list attributes evaluation score. The participant systems are ranked according to SF_{value}.

4 Performance of the Participants

In this bakeoff, 6 teams submitted 6 valid results. The team ID and the corresponding participants are listed in Table 4.

Team ID	Organization
CIST-BUPT	北京邮电大学
ICTNET_002	中国科学院计算所
WZ_v4	法国 INALCO
BLCU-yudong	北京语言大学
Result-BUPT	北京邮电大学
CASIA_CUC_PAES	中国科学院自动化所

Table 4. The Bakeoff Participants

The achieved performances of these systems under lenient and strict evaluations, are shown in Figure 2 and Figure 3, respectively. the performances of Personal Attributes Extraction in Chinese Text (the SF_Value) are uniformly lower than 0.5. Especially the ListScore lower than 0.4.

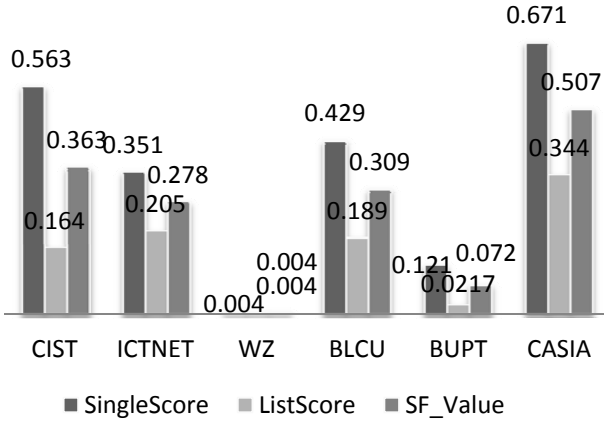

Figure 2. The lenient evaluation results

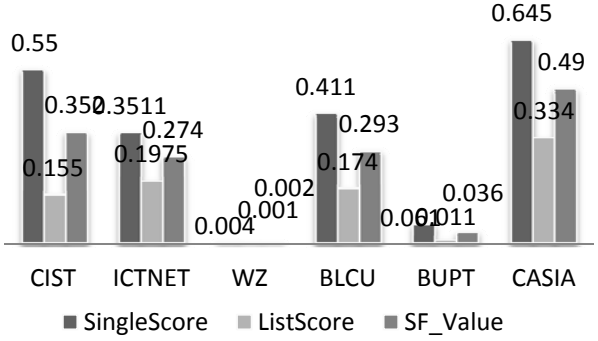

Figure 3. The strict evaluation results

Three participants submit the technical reports for this task.

Dong YU et al. [6] use a mixture framework consists of supervised learning and rule based extractor and human knowledge database. Firstly, they divide 25 attributes into several groups. A specific combination of methods for extracting the values for each group is developed. The CRF model and regular expression are employed to extract the instances, and the protagonist dependency relationship based filter and attribute keywords based filter are employed to post-process the answers extract. This system achieves the SF_Value of 0.309 under lenient evaluation and 0.293 under strict evaluation.

Kailun Zhang et al. [7] propose a method based on the combination of trigger words, dictionary and rules. This system narrow down the extraction scope by building attributes trigger words. The attributes such as state, province, and school, the cause of death and some similar fixed attributes are extracted by dictionary lookup directly through building the attributes dictionary. Some attributes extraction rules are developed to extract other instances. This system achieves the SF_Value of 0.363 under lenient evaluation and 0.352 under the strict evaluation.

Zhen Wang et al. [8] use a dependency patterns matching technique to extract the attribute instances. In order to get the ontology, they use some patterns to match dependency relations and save the extracted information into RDF format file. An alignment process is used to group same classes and remove duplicates in RDF files. Finally, they align their ontology to CLP's. The performance of this system may be limited to some language process problems. It achieves SF_Value of 0.0043 under lenient evaluation and 0.0025 under strict evaluation.

The top performance system, CASIA_CUC_PAES did not provide the technical report. This system achieves SF_Value of 0.507 under lenient evaluation and 0.490 under strict evaluation.

5 Analysis

The SF_Value performances of Personal Attributes Extraction in Chinese Text systems are lower than 0.5 while the Single Score is lower than 0.7 and the ListScore is lower than 0.4. In this section, we analyze the factors influence the extraction performance.

(1) One object sometimes have different expressions in Chinese language, for example,

the capital of China 北京 can be expressed as 北京市 or 京，and even the date 1990 年 5 月 6 日 can be expressed as May 6, 1941, or 1990-5-6, or 5/6/1990 and so on. The extraction system has the difficulty to extract all of these instances.

(2) In this evaluation, most system distinguish the titles and the alternate names hardly. Generally, alternate names refer to the assigned persons that are distinct from the "official" name. Alternate names may include aliases, stage names, alternate transliterations, abbreviations, alternate spellings, nicknames, or birth names. Compared with other slots, more inference should be used for selecting appropriate fillers for Alternate Names because the canonical names of entities often absent from source documents. As for the Titles or other extraneous information added to a name do not justify an alternate name. Generally, a given name alone is not a correct alternate name unless the person is unambiguously known that way.

(3) The administrative region divisions in different countries are not the same. Thus, most systems distinguish the city and the state or province hardly. For example, the 福冈县 in Japan is divided as state or province level, but the 浮山县 in China should be divided as city level. In the bakeoff, the geopolitical entities are divided to three levels (city, town, or village). Thus, these attributes are hardly distinguished, especially for the statistical-based system.

(4) Another problem is that attributes of string value are not be extracted exactly. For example, a mention of a serious illness is not an acceptable filler of cause of death unless it is explicitly linked to the death of the assigned person in the document. Assessors should be lenient in their judgment of the fullness of selected strings for cause of death. These types of attributes are basically a “catch all”, meaning that their fillers cannot be neatly classified as names or values. The text excerpts (or “strings”) that make up these fillers can sometimes be just a name, but are often expected to be more than a name.

Due to various factors and complication of the evaluations, the organizer may only ensure the relative fairness for each system. Meanwhile, it is observed that some errors in the submitted results are come at very small points. The carefully development will be helpful.

Furthermore, to make the evaluation results comparable, the organizer should use a uniform standard in te evaluation (besides the SingleScore, ListScore, and the SF_Value).

6 Conclusion

The Personal Attributes Extraction in Chinese Text task for CLP2014 has raised the problem in Chinese personal attributes extraction. Besides the basic difficulty of Chinese nature language processing and information extraction, there are other difficulties like common words detection, co-reference resolution. 6 teams have submitted their results. Most teams use rule-based methods or matching techniques while other team utilizes the statistical-based technique. Some proposed techniques are shown effective in person attribute extraction. The organizer expects this bakeoff is helpful to the research on person attribute extraction in Chinese text.

Acknowledgements

This study was supported by the National Natural Science Foundation of China No. 61370165, Natural Science Foundation of Guangdong Province S2013010014475, MOE Specialized Research Fund for the Doctoral Program of Higher Education 20122302120070, Open Projects Program of National Laboratory of Pattern Recognition, Shenzhen International Co-operation Research Funding GJHZ20120613 110641217, Shenzhen Foundational Research Funding JCYJ20120613152557576.

Reference

[1] Heng Ji and Raslfph Grishman. 2011. Knowledge Base Population: Successful Approaches and Challenges. Proc. 49th Annual Meeting Assn. Computational Linguistics.

[2] Zheng Chen, Suzanne Tamang, Adam Lee, Xiang Li, Wen-Pin Lin, Matthew Snover, Javier Artiles, MarissaPassantino and Heng Ji. 2010. CUNYBLENDER TAC-KBP2010 Entity Linking and Slot Filling System Description. Proc. TAC 2010 Workshop.

[3] Guillermo Garrido, Anselmo Peñas and Bernardo, Cabaleiro. 2013. UNED Slot Filling and Temporal Slot Filling systems at TAC KBP 2013. System description. Proc.

TAC 2013 Workshop.

[4] Bryan Kisiel, Justin Betteridge, Matt Gardner, Jayant Krishnamurthy, Ndapa Nakashole, Mehdi Samadi, Partha Talukdar, Derry Wijaya, Tom Mitchell. 2013. CMUML System for KBP 2013 Slot Filling. Proc. TAC 2013 Workshop.

[5] Joe Ellis, Heather Simpson, Kira Griffitt, Hoa Trang Dang etc, TAC KBP Slot. http://projects.ldc.upenn.edu/kbp/.

[6] Dong YU, Cheng YU, Gongbo TANG, Qin QU, Chunhua LIU, Yue TIAN, Jing YI. 2014. An Introduction to BLCU Personal Attributes Extraction System. Proc. Third SIGHAN Workshop on Chinese Language Processing.

[7] Kailun Zhang, Mingyin Wang, Xiaoyue Cong, Fang Huang, Hongfa Xue, Lei Li. 2014. Personal Attributes Extraction Based on the Combination Trigger Words, Dictionary and Rules. Proc. Third SIGHAN Workshop on Chinese Language Processing.

[8] Zhen Wang. 2014. Extraction system for Personal Attributes Extraction of CLP2014. Proc. The Third SIGHAN Workshop on Chinese Language Processing.