Chinese Personal Name Disambiguation Based on Person Modeling

Hua-Ping ZHANG[1] **Zhi-Hua LIU**[2] **Qian MO**[3] **He-Yan HUANG**[1]
[1] Beijing Institute of Technology, Beijing, P.R.C 100081
[2] North China University of Technology, P.R.C 100041
[3] Beijing Technology and Business University, Beijing, P.R.C 100048
Email: kevinzhang@bit.edu.cn

Abstract

This document presents the bakeoff results of Chinese personal name in the First CIPS-SIGHAN Joint Conference on Chinese Language Processing. The authors introduce the frame of person disambiguation system LJPD, which uses a new person model. LJPD was built in short time, and it is not given enough training and adjustment. Evaluation on LJPD shows that the precision is competitive, but the recall is very low. It has more space for further improvement.

1 Introduction

We participated in the First CIPS-SIGHAN Joint Conference on Chinese Language Processing. And have taken task 3: Chinese Personal Name disambiguation.

Chinese personal name disambiguation includes two stages: words are segmented to recognize Chinese personal name, and documents are clustered to disambiguate different person with the same personal name.

In our system, it involves the following steps:

1) Segmenting words and tagging the part-of-speech, and then recognizing Chinese personal name using ICTCLAS 2010 system[1].

2) Extracting personal feature to create the person attribution model on each document.

3) Generating initial clusters according to features in person model, and then clustering the initial clusters until the stop criteria is reached. The processing flow is illustrated in figure 1.

[1] It can be downloaded from http://hi.baidu.com/drkevinzhang

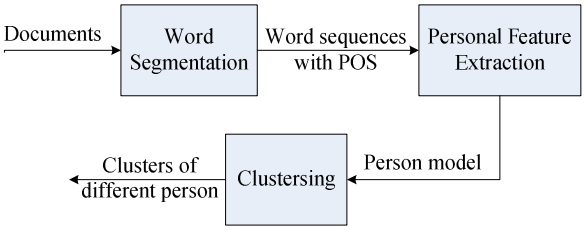

Figure 1 Step of Person Disambiguation

As illustrated in figure 1, the whole system addresses four problems: personal name recognition, anaphora resolution of personal name, person model creation and clustering.

2 Personal Name Recognition

Chinese personal name recognition is more difficult than English. Such difficulties usually combine with Chinese word segmentation. The set of Chinese personal name is infinite, and the rule of name construction is varied. Chinese personal name is often made up of a usual word, and has ambiguity with its context.

To solve the difficulties mentioned above, Chinese personal name recognition based on role tagging is given in [Zhang etc., 2002]. The approach is: tokens after segmentation are tagged using Viterbi algorithm with different roles according to their functions in the generation of Chinese personal name; the possible names are recognized after maximum pattern matching on the roles sequence [ZHANG, etc., 2002]. With this approach, the precision of ICTCLAS reaches 95.57% and the recall is 95.23% in an opening corpus which contains 1,108,049 words. In the corpus, the count of the personal name is 15,888. And ICTCLAS is a Chinese lexical analysis system witch combines part-of-speech

tagging, word segmentation, unknown words recognition. It can meet our requirements, so ICTCLAS provides personal name recognition in our system.

3 Anaphora Resolution of Personal Name

Anaphora is very common in natural language. Resolve this problem can help us get more information of the person from a document.

Anaphora resolution of personal name is an important part of anaphora resolution. At present, much advancement in anaphora resolution have occurred [Saliha 1998]. Anaphora resolution of personal pronouns is an especially complicate problem in anaphora resolution of personal name. In our system, we don't process this problem. The reason is that anaphora resolution of personal name will take side effect to personal name disambiguation unless its precision is definitely high. So we just process the anaphora of the first name or the second name. For example, "Jianmin Wang" in above context and "Professor Wang" will be resolved in our system.

4 Personal Model

We propose a person model to represent the person in the document:

$$\text{Person} = \{N, P, Q, R\}$$

where:

N is the collection of appellation of person, such as name, nickname, alias, and so on

P is the collection of the basic attributes of person

Q is the collection of the other attributes of person

R is the collection of the terms co-occurrence with person name, witch is called term field

In the system, we focused on seven attributes such as sex, nationality, birthday, native place, address, profession, family members and personal name, co-occurrence terms. In these features, name $\in$ N, {sex, nationality, birthday, native place} $\in$ P, {address, profession, family members} $\in$ Q, {co-occurrence term} $\in$ R. Table 1 is the examples of person model.

In view of the co-occurrence personal name is especially important for person disambiguation. We separate it as another field in R.

4.1 Attributes Feature

The components N, P and Q of person model are attributes feature. The dimension of these features for a person is different. For example, the sex of a person is constant in life, while his or her address may be different in different time. Take DOM to represent the dimension of the attributes features. Then:

$$DOM(N_i) = 1; (1 \leqslant i \leqslant n)$$

$$DOM(P_i) = 1; (1 \leqslant i \leqslant k)$$

$$DOM(Q_i) \geqslant 1; (1 \leqslant i \leqslant m)$$

For a person, N and P are constant in life. If one attribute of N or P between two persons is different, they are not the same person.

To get the attributes feature, we have three steps: First, segment word and tag part-of-speech for the input document. Second, we identify the triggering word which is defined as attributes value and the Max-Noun Phrase. The triggering words are identified by their POS and a hand-built triggering word thesaurus. At last, a classifier determines the attribute belongs to the left personal name or the right to the attribute. The classifier is trained by the corpus which is hand-tagged documents from internet.

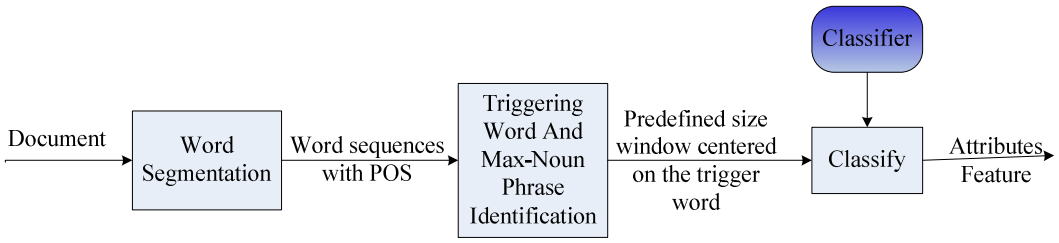

Figure 2 Step of Person Attributes Extraction

4.2 Term Field

In person model, R is the collection of the terms co-occurrence within person. We adopt Vector Space Model to represent these terms. The i-th term is represented by t_i, and its weight is represented by w_i, and the weight shows the importance of the term for the person.

$$R = (t_1, w_1; t_2, w_2; \ldots ; t_H, w_H)$$

To get the person's term field, we identify a scope witch these terms occurred. We consider three kinds of scope for term field: the total document, the paragraph where the personal name is present, sentence where the personal name is present. And then segment words and tag part-of-speech for these fragments. Next, filter out the attribute terms and filter by part-of-speech and leave only nouns, verb, adjective, adverb and name entry. Third, we make a stop word list, and filter out these stop terms. Last, according to the term's DF, filter out high frequency and low frequency terms, and only the terms witch DF is not lower than 2 and not higher than N/3(N is the total count of documents) are left.

In collection R, we have separated term field to co-occurrence personal name vector and co-occurrence common term vector. Because the two vectors have different affect to person disambiguation. This difference manifests in the different method to compute these weight. The common term's weight is computed by tf-idf algorithm:

$$w(t,\vec{d}) = \log(tf(t,\vec{d})+1)\times\log(N/n_t+1)$$

where:

$w(t,\vec{d})$ is the weight of term t in document $\vec{d}$

$tf(t,\vec{d})$ is the frequency of occurrence of t in $\vec{d}$

N is the total count of documents

n_t is the count of documents which contain term t

	sex	nationality	birthday	Native place	address	Family members	profession	Co-occurrence personal name	Co-occurrence terms field
Name1	男	汉	1949		北京		演员	…	……
Name2	女			山东		王红	教师	…	……
Name3	男	蒙			安徽		书记	…	……

Table 1 Examples of Person Model

The co-occurrence personal name's weight is computed below:

$$w(name,\vec{p}) = \log(nf(name,\vec{p})+1)\times\log(N'/n_{name}+1)$$

where:

$w(name,\vec{p})$ is the weight of co-occurrence name $name$

$nf(name,\vec{p})$ is the frequency of co-occurrence of $name$ and person $\vec{p}$

$name$ is the count of the co-occurrence of $name$ and the other personal name

The similarity of term field between two persons is calculated by the angle cosine:

$$Sim(X,Y) = \cos(X,Y) = \frac{\sum_i x_i * y_i}{\sqrt{\sum_i x_i^2 * \sum_i y_i^2}}$$

5 Clustering

Person model "Person = {N, P, Q, R}" is multi-dimensional. First, we adopted two rules to generate original clusters:

Rule 1: For two persons whose name is same, if one of the birthday (accurate to month) or relative is matched, these two persons are the same person.

Rule 2: For two persons whose name is same, if one of the sex, nationality, native place or birthday is not matched, these two persons are different.

There are profession, co-occurrence personal name and co-occurrence common terms left. For two persons whose name is same, we apply rule 1 and 2 first. If both of the two rules are not activating, compute the similarity $Sim_{position}(X, Y)$, $cos_{name}(X, Y)$, $cos_{term}(X, Y)$. And then synthesize these three similarities.

Assume the three factors profession, co-occurrence personal name and co-occurrence common terms are independent, and adopt Stanford certainty theory to synthesize the three similarities. The Stanford certainty theory creates confidence measures and some simple rules for combing these confidences. Assume E1, E2, E2 are the Stanford certainty factors of event B, and CF represent the confidence, then the confidence of event B is :

$$CF(B) = CF(E1) + CF(E2) + CF(E3) - CF(E1) \times CF(E2) - CF(E1) \times CF(E3) - CF(E2) \times CF(E3) + CF(E1) \times CF(E2) \times CF(E3)$$

For example, if the confidence of the three factors for event B is respectively: 88%, 74%, 66%, then the confidence for event B is $88\% + 74\% + 66\% - 88\% \times 74\% - 88\% \times 66\% - 76\% \times 66\% + 88\% \times 74\% \times 66\% = 98.93\%$.

To compute the confidence of the factors, we should get the threshold (represented by u_i) of the similarity for factors. If the similarity of the factor reaches the threshold, its confidence is 100%:

$$CF(E_i) = sim_{E_i} / u_i \qquad CF(E_i) \in [0,1]$$

The training method is: clustering training data according to the single factor, select the threshold with which the recall is higher with the premise that the precision is not lower than 98%. We get three thresholds 3, 0.5, 0.25 respectively for factor profession, co-occurrence personal name and co-occurrence common terms.

Overall, the algorithm takes two steps:

1) Adopt rule 1 and 2 to group the persons to the original clusters
2) Adopt agglomerative hierarchical clustering algorithm to clustering these original clusters.
   (1) Take each original cluster as a single cluster
   (2) Select two clusters which are most likelihood and merge to one cluster
   (3) If there is only one cluster or reaches stop criteria, exit. Else, go to step (2).

In the process of merging the clusters, we should merge the fragment of person. For term field vector, we simply compute the average of the term weights. For attribute feature, we adopt rule method to merge two clusters.

6 Task

We would introduce the operation of some different track in this section.

In formal test, we first get a query name and its all files. Then we segment these files and extract the related information of our person model and output to files. At last, we cluster these person models and output to result xml.

In the diagnosis test, the basic process is same to the formal test. The difference is that the element of clustering is changed to the subfolder of a real name. When all the subfolders are clustered for a query name, we merge the results to one xml file.

	B-Cubed			P-IP		
	precision	recall	F score	P	IP	F score
Formal test	80.2	68.75	68.4	86.12	76.37	77.54
Diagnosis test	94.62	63.32	72.48	96.44	72.78	80.85

Table 2 Evaluation result of Personal Disambiguation

7 Conclusion

Through the first bakeoff, we have learned much about the development in Chinese personal name recognition and person disambiguation. At the same time, we really find our problems during the evaluation. The bakeoff is interesting and helpful. We look forward to participate in forthcoming bakeoff.

References

ZHANG Hua-Ping, LIU Qun, YU Hong-Kui, CHENG Xue-Qi, BAI Shuo. *Chinese Named Entity Recognition Using Role Model*. International Journal of Computational Linguistics and Chinese language processing, 2003,Vol. 8 (2)

Azzam Saliha, Kevin Humphreys & Robert Gaizauskas. *Coreference resolution in a multilingual information extraction*. In the Proc. of the Workshop on Linguistic Coference. Granada, Spain.1998.

Yu Manquan. *Research on Knowledge Mining in Person Tracking*. Ph.D.Thesis of GUCAS. 2006