Chinese Personal Name Disambiguation:

Technical Report of Natural Language Processing Lab of Xiamen University

Xiang Zhu, Xiaodong Shi, Ningfeng Liu, YingMei Guo, Yidong Chen
Natural Language Processing Lab, Department of Cognitive Science, Xiamen University , Xiamen 361005
zhuxiang_sm@163.com,mandel@xmu.edu.cn,nffliu@gmail.com

Abstract

This report presents the work of our group in the Chinese personal name disambiguation workshop. We propose a system which uses a HAC algorithm to cluster the mentions referring to the same person with features extracted from the documents.

1 Introduction

Personal name disambiguation is actually a task to group those documents according to whether the given personal name appearing in that document refers to the same person in reality. It becomes an active research topic recently, and the evaluation campaign for the English personal name has been held twice.

Chinese personal name disambiguation is thought to be more challenging due to the need for word segmentation which could bring errors into the subsequent processes.

The most widely used method for personal name disambiguation is unsupervised clustering, we adopt a Hierarchical Agglomerative Clustering algorithm in our system, which is also the most popular clustering algorithm used by the teams of the second Web People Search Evaluation campaign.

The remaining part of this report is organized as follows. Section 2 introduces the document preprocessing. Section 3 describes the feature used for the clustering and the section 4 addresses the clustering algorithm. The workshop requests two different tests, a formal test, and a diagnosis test, we will discuss the difference between them with our system's result in section 5.

2 Document preprocessing

Different from the document preprocessing for English, we have to use a segmentation tool and a part-of-speech tagger to do some preprocessing work. For example, without the segmenting process, the documents only contain the string "最高军事法院" could be clustered when the query name is "高军", and we need the part-of-speech tagger to detect whether the Chinese word "黄海" stands for a personal name or a toponym.

Our experiments prove that the system is sensitive to the tools' performance, the different result between the formal test and the diagnosis test also proves this.

These tools are trained with the 90% data of The People's Daily published in Jan.1998,and tested with the others 10% data.

The performances of the tools are：

	Precision	Recall	F
seg	97.746%	97.793%	97.769%
tag	94.197%	94.242%	94.219%

Table 1: the performances of the tools

3 Extracted features

As mentioned previously, our goal is to group those documents. In our approach, each document is represented by a vector of features extracted from it automatically. We use three kinds of token-based features:

1) Nouns occur around the ambiguous personal name.

This kind of features is selected based on the idea that words around the ambiguous personal name are more relevant to it, and Nouns can provide a more diagnostic description of the person. We use a window to select the nouns.

2) Personal names (except the ambiguous personal name) and toponyms occur in the document.

It is intuitive that the identical person often associated with the same personal names and toponyms.

3) Words with high TFIDF value. In our final system, we use the ten words with highest TFIDF value.

This kind of features can reflect the theme of an article, an identical person often be mentioned in articles with the same theme.

Using these features simultaneously can alleviate the problem caused by spare data. The following table presents a quantitative analysis:

used features	highest F score on dev data
Feature 1	83.76
Feature 1,2	86.18
Feature 1,3	84.83
Feature 1,2,3	86.76

Table 2:analysis of features

These results are obtained by using a initial version of the preprocessing toools, and when we improve the performances of the tools, the highest F score increases from 86.76 to 89.61 .

The similarity between documents was measured with the cosine of feature vectors. When computing the similarity between two documents, we proposed a weighting method for the features as:

If the token occurs in the document, the weight will be 1, else 0.

We have tried another method that count the tokens' weight by its frequency, but experiments prove it is a less effective one, we can interpret it by this example:

If a word "教授" which refers to a person named "李明" appears twice in one document while three times in another document, these two documents are very likely referring to the same person, but the latter weighting method decreases the similarity between them.

4 Clustering algorithm

A Hierarchical Agglomerative Clustering algorithm is adopted in our system, which determines the number of the cluster by a fixed similarity threshold learned from the train data. At each stage of the clustering, the two most similar clusters are merged into a new one, and other clusters' similarities with the new cluster is the larger one of their similarities with the two old clusters.

We have tried some other clustering methods such as modified k-means clustering with the same features, but the performances are worse.

A pre-judgment stage before clustering is useful in the experiment, which can be done as follow:

If two documents have the same triple tokens "token1 token2 PersonName" (token2 is tagged as noun), then they are classified to one cluster.

5 Result and discussion

The difference between the formal test and the diagnosis test is that the ambiguous personal name in each document have been told in the latter, but you have to find it in the former by yourself. The method we adopt to detect the ambiguous personal name in the formal test is to find the token which is tagged as personal name while contains the query name.

Our system's performances are:

B-Cubed	precision	Recall	F
Formal test	90.55	84.88	85.72
Diagnosis test	89.84	89.84	89.08

Table 3: the performances in the B-Cubed criterion

P-IP	P	IP	F
Formal test	93.3	89.22	89.9
Diagnosis test	92.77	93.33	92.71

Table 4: the performances in the P-IP criterion

From the results we can know that Chinese personal name disambiguation can be affected by the segmentation tool and the part-of-speech tagger.

References

Y. Chen, S. Y. M. Lee, and C.-R. Huang. Polyuhk: A robust information extraction system for web personal names. In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.

J. Gong and D. Oard. Determine the entity number in hierarchical clustering for web personal name disambiguation. In 2nd Web

People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.

P. Kalmar and D. Freitag. Features for web person disambiguation. In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.