# A Pipeline Approach to Chinese Personal Name Disambiguation

**Yang Song, Zhengyan He, Chen Chen, Houfeng Wang**
Key Laboratory of Computational Linguistics (Peking University)
Ministry of Education,China
{ysong, hezhengyan, chenchen, wanghf}@pku.edu.cn

## Abstract

In this paper, we describe our system for Chinese personal name disambiguation task in the first CIPS-SIGHAN joint conference on Chinese Language Processing(CLP2010). We use a pipeline approach, in which preprocessing, unrelated documents discarding, Chinese personal name extension and document clustering are performed separately. Chinese personal name extension is the most important part of the system. It uses two additional dictionaries to extract full personal names in Chinese text. And then document clustering is performed under different personal names. Experimental results show that our system can achieve good performances.

## 1 Introduction

Personal name search is one of the most important tasks for search engines. When a personal name query is given to a search engine, a list of related documents will be shown. But not all of the returned documents refer to the same person whom users want to find. For example, the query name "jordan" is submitted to a search engine, we can get a lot of documents containing "jordan". Some of them may refer to the computer scientist, others perhaps refer to the basketball player. For English, there have been three Web People Search (WePS[1]) evaluation campaigns on personal name disambiguation. But for Chinese, this is the first time. It encounters more challenge for Chinese personal name disambiguation. There are no word boundary in Chinese text, so it becomes difficult to recognize the full personal names from Chinese text. For example, a query name "高明" is given, but the full personal name from some documents may be an extension of "高明", like "高明光" or "高明珍", and so on. Meanwhile, "高明" can also be a common Chinese word. So we need to discard those documents which are not refered to any person related to the given query name.

To solve the above-mentioned problem, we explore a pipeline approach to Chinese personal name disambiguation. The overview of our system is illustrated in Figure 1. We split this task into four parts: preprocessing, unrelated documents discarding, Chinese personal name extension and document clustering. In preprocessing and unrelated documents discarding, we use word segmentation and part-of-speech tagging tools to process the given dataset and documents are discarded when the given query name is not tagged as a personal name or part of a personal name. After that we perform personal name extension in the documents for a given query name. When the query name has only two characters. We extend it to the left or right for one character. For example, we can extend "林鹏" to "金林鹏" or "林鹏飞". The purpose of extending the query name is to obtain the full personal name. In this way, we can get a lot of full personal names for a given query name from the documents. And then document clustering

[1] http://nlp.uned.es/weps/

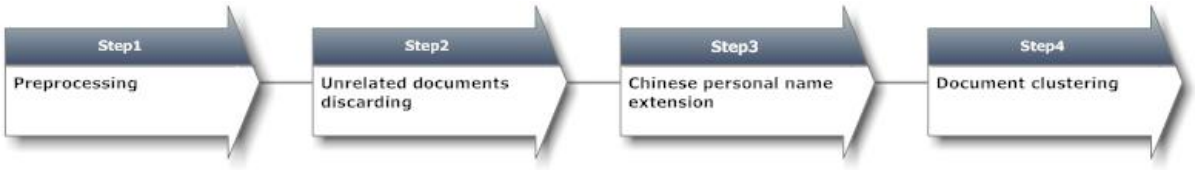


Figure 1: Overview of the System

is performed under different personal names. HAC (Hierarchical Agglomerative Clustering) is selected here. We represent documents with bag of words and solve the problem in vector space model, nouns, verbs, bigrams of nouns or verbs and named entities are selected as features. The feature weight value takes 0 or 1. In HAC, we use group-average link method as the distance measure and consine similarity as the similarity computing measure. The stopping criteria is dependent on a threshold which is obtained from training data. Our system produces pretty good results in the final evaluation.

The remainder of this paper is organized as follows. Section 2 introduces related work. Section 3 gives a detailed description about our pipeline approach. It includes preprocessing, unrelated documents discarding, Chinese personal name extension and document clustering. Section 4 presents the experimental results. The conclusions are given in Section 5.

## 2 Related Work

Several important studies have tried to solve the task introduced in the previous section. Most of them treated it as an clustering problem. Bagga & Baldwin (1998) first selected tokens from local context as features to perform intra-document coreference resolution. Mann & Yarowsky (2003) extracted local biographical information as features. Niu *et al.* (2004) used relation extraction results in addition to local context features and get a perfect results. Al-Kamha and Embley (2004) clustered search results with feature set including attributes, links and page similarities.

In recent years, this problem has attracted a great deal of attention from many research institutes. Ying Chen *et al.* (2009) used a Web 1T 5-gram corpus released by Google to extract additional features for clustering. Masaki Ikeda *et al.* (2009) proposed a two-stage clustering algorithm to improve the low recall values. In the first stage, some reliable features (like named entities) are used to connect documents about the same person. After that, the connected documents (document cluster) are used as a source from which new features (compound keyword features) are extracted. These new features are used in the second stage to make additional connections between documents. Their approach is to improve clusters step by step, where each step refines clusters conservatively. Han & Zhao (2009) presented a system named CASIANED to disambiguate personal names based on professional categorization. They first categorize different personal name appearances into a real world professional taxonomy, and then the personal name appearances are clustered into a single cluster. Chen Chen *et al.* (2009) explored a novel feature weight computing method in clustering. It is based on the pointwise mutual information between the ambiguous name and features. In their paper, they also develop a trade-off point based cluster stopping criteria which find the trade-off point between intra-cluster compactness and inter-cluster separation.

Our approach is based on Chinese personal name extension. We recognize the full personal names in Chinese text and perform document clustering under different personal names.

## 3 Methodology

In this section, we will explain preprocessing, unrelated documents discarding, Chinese

personal name extension and document clustering in order.

### 3.1 Preprocessing

We use ltp-service[2] to process the given Chinese personal name disambiguation dataset (a detailed introduction to it will be given in section 4). Training data in the dataset contains 32 query names. There are 100-300 documents under every query name. All the documents are collected from Xinhua News Agency. They contain the exact same string as query names. Ltp-service is a web service interface for LTP[3](Language Technology Platform). LTP has integrated many Chinese processing modules, including word segmentation, part-of-speech tagging, named entity recognition, word sense disambiguation, and so on. Jun Lang *et al.* (2006) give a detailed introduction to LTP. Here we only use LTP to generate word segmentation, part-of-speech tagging and named entity recognition results for the given dataset.

### 3.2 Unrelated documents discarding

Under every query name, there are 100-300 documents. But not all of them are really related. For example, "高军" is a query name in training data. In corresponding documents, some are refered to real personal names like "高军" or "高军田". But others may be a substring of an expression such as "最高军事法院". These documents are needed to be filtered out. We use the preprocessing tool LTP to slove this problem. LTP can do word segmentation and part-of-speech tagging for us. For each document under a given query name, if the query name in the document is tagged as a personal name or part of some extended personal name, the document will be marked as undiscarded, otherwise the document will be discarded. Generally speaking, for the query name containing three characters, we don't need to discard any of the corresponding documents. But in practice, we find that for some query names, LTP always gives the invariable part-of-speech. For example, no matter what the context of "黄海" is, it is always tagged as a geographic name. So we use another preprocessing tool ICTCLAS[4]. Only when both of them mark one document as discarded, we discard the corresponding document.

### 3.3 Chinese personal name extension

After discarding unrelated documents, we need to recognize the full Chinese personal names. We hypothesize that the full Chinese personal name has not more than three characters (We don't consider the compound surnames here). So the query names containing only two Chinese characters are considered to extend. In our approach, we use two Chinese personal names dictionaries. One is a surname dictionary containing 423 one-character entries. We use it to do left extend for the query name. For example, the query name is "高明" and its left character in a document is "刘", we will extend it to full personal name "刘高明". The other is a non-ending Chinese character dictionary containing 64 characters which could not occur at the end of personal names. It is constructed by a personal title dictionary. We use every title's first character and some other special characters (such as numbers or punctuations) to constuct the dictionary. Some manual work has also been done to filter a few incorrect characters. Several examples of the two dictionaries are shown in Table 1.

Through the analysis of Xinhua News articles, we also find that nearly half of the documents under given query name actually refer to the reporters. And they often appear in the first or last brackets in the body of corresponding document. For example, "(通讯员刘国党、黄海生)" is a sentence containing query name "黄海". We use some simple but efficient rules to get full personal names for this case.

### 3.4 Document clustering

For every query name, we can get a list of full peronal names. For example, when the

[2]http://code.google.com/p/ltp-service/

[3]http://ir.hit.edu.cn/ltp/

[4]http://ictclas.org/

Table 1: Several Examples of the two Dictionaries

| Dictionaries | Examples |
|---|---|
| Surnames | 王, 张, 李, 陈, 刘, 杨, 黄, 吴, 周, 赵, 徐, 孙, 朱, 胡... |
| Non-ending Chinese characters | 说, 的, 县, 市, 坐, 反, 讲, 跳, 牌, 各, 住, 在, 仍, 打... |

query name is "郭勇", we can get the personal names like "郭勇民", "郭勇军", "郭勇勤", "郭勇孝". And then document clustering is performed under different personal names.

#### 3.4.1 Features

We use bag of words to represent documents. Some representative words need to be chosen as features. LTP can give us POS tagging and NER results. We select all the nouns, verbs and named entities which appear in the same paragraph with given query name as features. Meanwhile, the bigrams of nouns or verbs are also selected. We take 0 or 1 for feature weight value. 0 represents that the feature doesn't appear in corresponding paragraphs, and 1 represents just the opposite. We find that this weighting scheme is more effective than TFIDF.

#### 3.4.2 Clustering

All features are represented in vector space model. Every document is modeled as a vertex in the vector space. So every document can be seen as a feature vector. Before clustering, the similarity between documents is computed by cosine value of the angle between feature vectors. We use HAC to do document clustering. It is a bottom-up algorithm which treats each document as a singleton cluster at the outset and then successively merges (or agglomerates) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. From our experience, single link and group-average link method seem to work better than complete link one. We use group-average link method in the final submission. The stopping criteria is a difficult problem for clustering. Here we use a threshold for terminating condition. So it is not necessary to determine the number of clusters beforehand. We select a threshold which produces the best performance in training data.

## 4 Experimental Results

The dataset for Chinese personal name disambiguation task contains training data and testing data. The training data contains 32 query names. Every query name folder contains 100-300 news articles. Given the query name, all the documents are retrived by character-based matching from a collection of Xinhua news documents in a time span of fourteen years. The testing data contains 25 query names. Two threshold values as terminating conditions are obtained from training data. They are 0.4 and 0.5. For evaluation, we use P-IP score and B-cubed score (Bagga and Baldwin, 1998). Table 2 & Table 3 show the official evaluation results.

Table 2: Official Results for P-IP score

| **Threshold** | **P-IP** | | |
|---|---|---|---|
| | **P** | **IP** | **F score** |
| 0.4 | 88.32 | 94.9 | 91.15 |
| 0.5 | 91.3 | 91.77 | 91.18 |

Table 3: Official Results for B-Cubed score

| **Threshold** | **B-Cubed** | | |
|---|---|---|---|
| | **Precision** | **Recall** | **F score** |
| 0.4 | 83.68 | 92.23 | 86.94 |
| 0.5 | 87.87 | 87.49 | 86.84 |

Besides the formal evaluation, the organizer also provide a diagnosis test designed to explore the relationship between Chinese word segmentation and personal name disambiguation. That means the query names in the documents are segmented correctly by manual work. Table 4 & Table 5 show the diagnosis results.

Table 4: Diagnosis Results for P-IP score

| Threshold | P-IP | | |
|---|---|---|---|
| | P | IP | F score |
| 0.4 | 89.01 | 95.83 | 91.96 |
| 0.5 | 91.85 | 92.68 | 91.96 |

Table 5: Diagnosis Results for B-Cubed score

| Threshold | B-Cubed | | |
|---|---|---|---|
| | Precision | Recall | F score |
| 0.4 | 84.53 | 93.42 | 87.96 |
| 0.5 | 88.59 | 88.59 | 87.8 |

The official results show that our method performs pretty good. The diagnosis results show that correct word segmentation can improve the evaluation results. But the improvement is rather limited. That is mainly because Chinese personal name extension is done well in our approach. So the diagnosis results don't gain much profit from query names' correct segmentation.

## 5 Conclusions

We describe our framework in this paper. First, we use LTP to do preprocessing for original dataset which comes from Xinhua news articles. LTP can produce good results for Chinese text processing. And then we use two additional dictionaries(one is Chinese surname dictionary, the other is Non-ending Chinese character dictionary) to do Chinese personal name extension. After that we perform document clustering under different personal names. Official evaluation results show that our method can achieve good performances.

In the future, we will attempt to use other features to represent corresponding persons in the documents. We will also investigate automatic terminating condition.

## 6 Acknowledgments

This research is supported by National Natural Science Foundation of Chinese (No.60973053) and Research Fund for the Doctoral Program of Higher Education of China (No.20090001110047).

## References

J. Artiles, J. Gonzalo, and S. Sekine. 2009. *WePS 2 evaluation campaign: overview of the web people search clustering task.* In 2nd Web People Search Evaluation Workshop(WePS 2009), 18th WWW Conference.

Bagga and B. Baldwin. 1998. *Entity-based cross-document coreferencing using the vector space model.* In Proceedings of 17th International Conference on Computational Linguistics, 79–85.

Mann G. and D. Yarowsky. 2003. *Unsupervised personal name disambiguation.* In Proceedings of CoNLL-2003, 33–40, Edmonton, Canada.

C. Niu, W. Li, and R. K. Srihari. 2004. *Weakly Supervised Learning for Cross-document Person Name Disambiguation Supported by Information Extraction.* In Proceedings of ACL 2004.

Al-Kamha. R. and D. W. Embley. 2004. *Grouping search-engine returned citations for person-name queries.* In Proceedings of WIDM 2004, 96-103, Washington, DC, USA.

Ying Chen, Sophia Yat Mei Lee, and Chu-Ren Huang. 2009. *PolyUHK:A Robust Information Extraction System for Web Personal Names.* In 2nd Web People Search Evaluation Workshop(WePS 2009), 18th WWW Conference.

Masaki Ikeda, Shingo Ono, Issei Sato, Minoru Yoshida, and Hiroshi Nakagawa. 2009. *Person Name Disambiguation on the Web by Two-Stage Clustering.* In 2nd Web People Search Evaluation Workshop(WePS 2009), 18th WWW Conference.

Xianpei Han and Jun Zhao. 2009. *CASIANED: Web Personal Name Disambiguation Based on Professional Categorization.* In 2nd Web People Search Evaluation Workshop(WePS 2009), 18th WWW Conference.

Chen Chen, Junfeng Hu, and Houfeng Wang. 2009. *Clustering technique in multi-document personal name disambiguation.* In Proceedings of the ACL-IJCNLP 2009 Student Research Workshop, pages 88–95.

Jun Lang, Ting Liu, Huipeng Zhang and Sheng Li. 2006. *LTP: Language Technology Platform.* In Proceedings of SWCL 2006.

Bagga, Amit and B. Baldwin. 1998. *Algorithms for scoring co-reference chains.* In Proceedings of the First International Conference on Language Resources and Evaluation Workshop on Linguistic co-reference.