A Multi-stage Clustering Framework for Chinese Personal Name Disambiguation

Huizhen Wang, Haibo Ding, Yingchao Shi, Ji Ma, Xiao Zhou, Jingbo Zhu
Natural Language Processing Laboratory,
Northeastern University
Shenyang, Liaoning, China
{wanghuizhen|zhujingbo@mail.neu.edu.cn
{dinghb|shiyc|maji}@mail.neu.edu.cn

Abstract

This paper presents our systems for the participation of Chinese Personal Name Disambiguation task in the CIPS-SIGHAN 2010. We submitted two different systems for this task, and both of them all achieve the best performance. This paper introduces the multi-stage clustering framework and some key techniques used in our systems, and demonstrates experimental results on evaluation data. Finally, we further discuss some interesting issues found during the development of the system.

1 Introduction

Personal name disambiguation (PND) is very important for web search and potentially other natural language applications such as question answering. CIPS-SIGHAN bakeoffs provide a platform to evaluate the effectiveness of various methods on Chinese PND task.

Different from English PND, word segmentation techniques are needed for Chinese PND tasks. In practice, person names are highly ambiguous because different people may have the same name, and the same name can be written in different ways. It's an n-to-n mapping of person names to the specific people. There are two main challenges on Chinese PND: the first one is how to correctly recognize personal names in the text, and the other is how to distinguish different persons who have the same name. For address these challenges, we designed a rule-based combination technique to improve NER performance and propose a multi-stage clustering framework for Chinese PND. We participated in the bakeoff of the Chinese PND task, on the test set and the diagnosis test set, our two systems are ranked at the 1st and 2nd position.

The rest of this paper is organized as follows. In Section 2, we first give the key features and techniques used in our two systems. In Section 3, experimental results on the evaluation test data demonstrated that our methods are effective to disambiguate the personal name, and discussions on some issues we found during the development of the system are given. In Section 4, we conclude our work.

2 System Description

In this section, we describe the framework of our systems in more detail, involving data preprocessing, *discard*-class document identification, feature definition, clustering algorithms, and sub-system combination.

2.1 Data Preprocessing

There are around 100-300 news articles per personal name in the evaluation corpus. Each article is stored in the form of XML and encoded in UTF-8. At first, each news article should be preprocessed as follows:

- Use a publicly available Chinese encoding Converter tool to convert each news article from UTF-8 coding into GB[1];
- Remove all XML tags;
- Process Chinese word segmentation, part-of-speech (POS) tagging and name entity recognition (NER);

The performance of word segmentation and NER tools generally affect the effectiveness of our Chinese PND systems. During system de-

[1] http://www.mandarintools.com/

veloping process, we found that the publicly available NER systems obtain unsatisfactory performance on evaluation data. To address this challenge, we propose a new rule-based combination technique to improve NER performance. In our combination framework, two different NER systems are utilized, including a CRF-based NER system and our laboratory's NER system (Yao et al.,2002). The latter was implemented based on the maximum matching principle and some linguistic post-preprocessing rules. Since both two NER systems adopt different technical frameworks, it is possible to achieve a better performance by means of system combination techniques.

The basic idea of our combination method is to first simply combine the results produced by both NER systems, and further utilize some heuristic post-processing rules to refine NE identification results. To achieve this goal, we first investigate error types caused by both NER systems, and design some post-preprocessing rules to correct errors or select the appropriate NER results from disagreements. Notice that such rules are learned from sample data (i.e., training set), not from test set. Experimental results demonstrate satisfactory NER performance by introducing these heuristic refinement rules as follows:

- **Conjunction Rules**. Two NEs separated by a conjunction (such as "和","或","与", "、") belong to the same type, e.g., "高明/adj.和/吴倩莲/person". Such a conjunction rule can help NER systems make a consistent prediction on both NEs, e.g., "高明/person" and "高峰/person".
- **Professional Title Rules**. Professional title words such as "主任" are strong indicators of person names, e.g., "主任/李刚". Such a rule can be written in the form of "*professional_title+person_name*".
- **Suffix Rules**. If an identified person name is followed by a suffix of another type of named entities such as location, it is not a true person name, for example, "阿萨姆邦德马杰/person 镇/的/居民". Since "镇" is a suffix of a location name. "阿萨姆邦德马杰/person 镇/location-suffix" should be revised to be a new location name, namely "阿萨姆邦德马杰镇/location".
- **Foreign Person Name Rules**. Two identified person names connected by a dot are merged into a single foreign person name, e.g., "菲/. /罗杰斯" => "菲. 罗杰斯"
- **Chinese Surname Rules**. Surnames are very important for Chinese person name identification. However, some common surnames can be single words depending upon the context, for example, the Chinese word "张" can be either a surname or a quantifier. To tackle this problem, some post-processing rules for "张, 段, 高, 刘, 赵" are designed in our system.
- **Query-Dependent Rules.** Given a query person name *A*, if the string *AB* occurring in the current document has been identified as a single person name many times in other documents, our system would tend to segment *AB* as a single person name rather than as *A/B*. For example, if "郭伟明" was identified as a true person name more than one time in other documents, in such a case, "议员/郭伟/明/表示/"=> "议员/郭伟明/person 表示/"

Incorporating these above post-processing rules, our NER system based on heuristic post-processing rules shows 98.89% precision of NER on training set.

2.2 Discard-Class Document Identification

Seen from evaluation data, there are a lot of documents belonging to a specific class, referred to as *discard*-class. In the discard-class, the query person name occurring in the document is not a true person name. For example, a query word "黄海" is a famous ocean name not a person name in the sentence "三江飘流分别可达日本海、**黄海**和鄂霍茨克海". In such a case, the corresponding document is considered as discard-class. Along this line, actually the discard-class document identification is very simple task. If a document does not contain a true person name that is the same as the query or contains the query, it is a discard-class document.

2.3 Feature Definition

To identify different types of person name and for the PND purpose, some effective binary fea-

tures are defined to construct the document representation as feature vectors as follows:

- **Personal attributes**: involving professional title, affiliation, location, co-occurrence person name and organization related to the given query.
- **NE-type Features**: collecting all NEs occurring in the context of the given query. There are two kinds of NE-type features used in our systems, local features and global features. The global features are defined with respect to the whole document while the local features are extracted only from the two or three adjacent sentences for the given query.
- **BOW-type features**: constructing the context feature vector based on bag-of-word model. Similarly, there are local and global BOW-type features with respect to the context considered.

2.4 A Multi-stage Clustering Framework

Seen from the training set, 36% of person names indicate journalists, 10% are sportsmen, and the remaining are common person names. Based on such observations, it is necessary to utilize different methodology to PND on different types of person names, for example, because the most effective information to distinguish different journalists are the reports' location and colleagues, instead of the whole document content. To achieve a satisfactory PND performance, in our system we design three different modules for analyzing journalist, sportsman and common person name, respectively.

2.4.1 PND on the Journalist Class

In our system, some regular expressions are designed to determine whether a person name is a journalist or not. For example:

- 新华社/ni */ns */t */t 消息|电/n (/w .* [《/w */ni 》/w]* query name/nh .*)/w
- (/w .*query name/nh .*)/w
- [*/nh]* query name/nh [*/nh]
- 记 者 | 编 辑 /n [*/nh]* query name/nh [*/nh]*

To disambiguate on the journalist class, our system utilizes a rule-based clustering technique distinguish different journalists. For each document containing the query person name as journalists, we first extract the organization and the location occurring in the local context of the query. Two such documents can be put into the same cluster if they contain the same organization or location names, otherwise not. In our system, a location dictionary containing province-city information extracted from Wikipedia is used to identify location name. For example: 辽宁省 (沈阳 大连 铁岭 鞍山 …), 河北(石家庄 唐山 秦皇岛 邯郸 邢台…). Based on this dictionary, it is very easy to map a city to its corresponding province.

2.4.2 PND on the Sportsman Class

Like done in PND on the journalist class, we also use rule-based clustering techniques for disambiguating sportsman class. The major difference is to utilize topic features for PND on the sportsman class. If the topic of the given document is sports, this document can be considered as sportsman class. The key is to how to automatically identify the topic of the document containing the query. To address this challenge, we adopt a domain knowledge based technique for document topic identification. The basic idea is to utilize a domain knowledge dictionary NEUKD developed by our lab, which contains more than 600,000 domain associated terms and the corresponding domain features. Some domain associated terms defined in NEUKD are shown in Table 1.

Domain associated term	Domain feature concept
足球队(football team)	Football, Sports
自行车队 (cycling team)	Traffic, Sports, cycling
中国象棋 (Chinese chess)	Sports, Chinese chess
执白(white side)	Sports, the game of go
芝加哥公牛 (Chicago bulls)	Sports, basketball

Table 1: Six examples defined in the NEUKD

In the domain knowledge based topic identification algorithm, all domain associated terms occurring in the given document are first mapped into domain features such as *football*, *basketball* or *cycling*. The most frequent domain feature is considered as the most likely topic. See Zhu and Chen (2005) for details.

Two documents with the same topic can be grouped into the same cluster.

Person name	Document no.	sports
杨波	081	篮球
杨波	094	射箭
杨波	098	射箭
杨波	100	射箭

Table 2: Examples of PND on Sportsman Class

2.4.3 Multi-Stage Clustering Framework

We proposed a multi-stage clustering framework for PND on common person name class, as shown In Figure 1.

In the multi-stage clustering framework, the first-stage is to adopt strict rule-based hard clustering algorithm using the feature set of personal attributes. The second-stage is to implement constrained hierarchical agglomerative clustering using NE-type local features. The third-stage is to design hierarchical agglomerative clustering using BOW-type global features. By combining those above techniques, we submitted the first system named NEU_1.

2.4.4 The second system

Besides, we also submitted another PND system named NEU_2 by using the single-link hierarchical agglomerative clustering algorithm in which the distance of two clusters is the cosine similarity of their most similar members (Masaki et al., 2009, Duda et al., 2004). The difference between our two submission systems NEU_1 and NEU_2 is the feature weighting method. The motivation of feature weighting method used in NEU_2 is to assume that words surrounding the query person name in the given document are more important features than those far away from it, and person name and location names occurring in the context are more discriminative features than common words for PND purpose. Along this line, in the feature weighting scheme used in NEU_2, for each feature extracted from the sentence containing the query person name, the weight of a word-type feature with the POS of “ns”, ”ni” or ”nh ” is assigned as 3, Otherwise 1.5; For the features extracted from other sentences, the weight of a word with the POS of “ns”or ”nh ” is set to be 2, the ones of “ni” POS is set to 1.5, otherwise 1.0.

Algorithm 1: Multi-stage Clustering Framework

Input: a person name *pn*, and its related document set D={d_1, d_2, …, d_m} in which each document d_i contains the person name *pn*;

Output: clustering results C={C_1,C_2, …,C_n}, where $\bigcup_i C_i = C$ and $C_i \cap C_j = \Phi$

For each $d_i \in D$ do
  $S_i = \{s|pn \in s, s \in d_i\}$;
  $ORG_i = \{t|t \in s, s \in S_i, POS(t) = ni\}$;
  $PER_i = \{t|t \in s, s \in S_i, POS(t) = nh\}$;
  $L_{di} = \{t|t \in s, s \in S_i\}$; //local feature set
  $G_{di} = \{t|t \in d_i\}$; //global feature set
  $C_i = \{d_i\}$;
End for

Stage 1: Strict rules-based clustering
Begin
  For each $C_i \in C$ do
    If $ORG_i \cap ORG_j \neq \Phi$ or
      $|PER_i \cap PER_j| \geq 2$
    Then $C_i = C_i \cup C_j$;
      $ORG_i = ORG_i \cup ORG_j$;
      $PER_i = PER_i \cup PER_j$;
      Remove C_j from C ;
  End for
End

Stage 2: Constrained hierarchical agglomerative clustering algorithm using local features
Begin
  Set each $c \in C$ as an initial cluster;
  do
    $[C_i, C_j] = \arg\max_{C_i, C_j \in C} sim(C_i, C_j)$
    $sim(C_i, C_j) = \max_{d_x \in C_i, d_y \in C_j} sim(d_x, d_y)$
      $= \max_{d_x \in C_i, d_y \in C_j} \cos(L_{d_x}, L_{d_y})$
    $C_i = C_i \cup C_j$;
    Remove C_j from C ;
  until $sim(C_i, C_j) < \theta$.
End

Stage 3: Constrained hierarchical agglomerative clustering algorithm using global features, i.e., utilize the same algorithm used in stage 2 by considering the global feature set G for cosine-based similarity calculation instead of the local feature set L.

Figure 1: Multi-stage Clustering Framework

2.5 Final Result Generation

As discussed above, there are many modules for PND on Chinese person name. In our NEU_1, the final results are produced by combining outputs of discard-class document clustering, journalist-class clustering, sportsman-class clustering and multi-stage clustering modules. In NEU-2 system, the outputs of discard-class document clustering, journalist-class clustering, sportsman-class clustering and single-link clustering modules are combined to generate the final results.

3 Evaluation

3.1 Experimental Settings

- Training data: containing about 30 Chinese person names, and a set of about 100-300 news articles are provided for each person name.
- Test data: similar to the training data, and containing 26 unseen Chinese personal names, provided by the SIGHAN organizer.
- Performance evaluation metrics (Artiles et al., 2009): B_Cubed and P_IP metrics.

3.2 Results

Table 3 shows the performance of our two submission systems NEU_1 and NEU_2 on the test set of Sighan2010 Chinese personal name disambiguation task.

System No.	B_Cubed			P_IP		
	P	R	F	P	IP	F
NEU_1	**95.76**	88.37	**91.47**	**96.99**	92.58	**94.56**
NEU_2	95.08	88.62	91.15	96.73	92.73	94.46

Table 3: Results on the test data

NEU-1 system was implemented by the multi-stage clustering framework that uses single-link clustering method. In this framework, there are two threshold parameters θ and μ. Both threshold parameters are tuned from training data sets.

After the formal evaluation, the organizer provided a diagnosis test designed to explore the relationship between Chinese word segmentation and personal name disambiguation. In the diagnosis test, the personal name disambiguation task was simplified and limited to the documents in which the personal name is tagged correctly. The performance of our two systems on the diagnosis test set of Sighan2010 Chinese personal name disambiguation task are shown in Table 4.

System no.	B_Cubed			P_IP		
	P	R	F	P	IP	F
NEU_1	**95.6**	89.74	**92.14**	**96.83**	93.62	**95.03**
NEU_2	94.53	89.99	91.66	96.41	93.8	94.9

Table 4: Results of the diagnosis test on test data

As shown in the Table 3 and Table 4, NEU-1 system achieves the highest precision and F values on the test data and the diagnosis test data.

3.3 Discussion

We propose a multi-stage clustering framework for Chinese personal name disambiguation. The evaluation results demonstrate that the features and key techniques our systems adopt are effective. Our systems achieve the best performance in this competition. However, our recall values are not unsatisfactory. In such a case, there is still much room for improvement. Observed from experimental results, some interesting issues are worth being discussed and addressed in our future work as follows:

(1) For PND on some personal names, the document topic information seems not effective. For example, the personal name "郭华(Guo Hua)" in training set represent one shooter and one billiards player. The PND system based on traditional clustering method can not effectively work in such a case due to the same sports topic. To solve this problem, one solution is to sufficiently combine the personal attributes and document topic information for PND on this person name.

(2) For the journalist-class personal names, global BOW-type features are not effective in this case as different persons can report on the same or similar events. For example, there are four different journalists named "朱建军(Zhu Jianjun)" in the training set, involving different locations such as Beijing, Zhengzhou, Xining or Guangzhou. We can distinguish them in terms of the location they are working in.

(3) We found that some documents in the training set only contain lists of news title and the news reporter. In this case, we can not discriminate the persons with respect to the location of entire news. It's worth studying some effective solution to address this challenge in our future work.

(4) Seen from the experimental results, some personal names such as "李刚(Li gang)" are wrong identified because this person is associated with multiple professional titles and affiliates. In this case, the use of exact matching methods can not yield satisfactory results. For example, the query name "李刚(Li gang)" in the documents 274 and 275 is the president of "中国对外文化交流协会(China International Culture Association)" while in the documents 202, 225 and 228, he is the director of "文化部对外文化联络局(Bureau of External Cultural Relations of Chinese Ministry of Culture)". To group both cases into the same cluster, it's worth mining the relations and underlying semantic relations between entities to achieve this goal.

4 Conclusion

This paper presents our two Chinese personal name disambiguation systems in which various constrained hierarchical agglomerative clustering algorithms using local or global features are adopted. The bakeoff results show that our systems achieve the best performance. In the future, we will pay more attention on the personal attribute extraction and unsupervised learning approaches for Chinese personal name disambiguation.

5 Acknowledgements

This work was supported in part by the National Science Foundation of China (60873091) and the Fundamental Research Funds for the Central Universities.

References

Artiles, Javier, Julio Gonzalo and Satoshi 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.

Duda, Richard O., Peter E.Hart, and David G.Stork. 2004. Pattern Classification. China Machine Press.

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

Yao, Tianshun, Zhu Jingbo , Zhang Li, Yang Ying. Nov. 2002. Natural Language Processing , Second Edition, Tsinghua press.

Zhu, Jingbo and Wenliang Chen. 2005. Some Studies on Chinese Domain Knowledge Dictionary and Its Application to Text Classification. In Proc. of SIGHAN4.