Combine Person Name and Person Identity Recognition and Document Clustering for Chinese Person Name Disambiguation

Ruifeng Xu[1,2],Jun Xu[1],Xiangying Dai[1]
Harbin Institute of Technology,
Shenzhen Postgraduate School, China
{xuruifeng.hitsz;hit.xujun;
mi-chealdai}@gmail.com

Chunyu Kit[2]
[2]City University of Hong Kong,
Hong Kong, China
ctckit@cityu.edu.hk

Abstract

This paper presents the HITSZ_CITYU system in the CIPS-SIGHAN bakeoff 2010 Task 3, Chinese person name disambiguation. This system incorporates person name string recognition, person identity string recognition and an agglomerative hierarchical clustering for grouping the documents to each identical person. Firstly, for the given name index string, three segmentors are applied to segment the sentences having the index string into Chinese words, respectively. Their outputs are compared and analyzed. An unsupervised clustering is applied here to help the personal name recognition. The document set is then divided into subsets according to each recognized person name string. Next, the system identifies/extracts the person identity string from the sentences based on lexicon and heuristic rules. By incorporating the recognized person identity string, person name, organization name and contextual content words as features, an agglomerative hierarchical clustering is applied to group the similar documents in the document subsets to obtain the final person name disambiguation results. Evaluations show that the proposed system, which incorporates extraction and clustering technique, achieves encouraging recall and good overall performance.

1 Introduction

Many people may have the same name which leads to lots of ambiguities in text, especially for some common person names. This problem puzzles many information retrieval and natural language processing tasks. The person name ambiguity problem becomes more serious in Chinese text. Firstly, Chinese names normally consist of two to four characters. It means that for a two-character person name, it has only one character as surname to distinguish from other person names with the same family name. It leads to thousands of people have the same common name, such 王刚 and 李明. Secondly, some three-character or four-character person name may have one two-character person name as its substring such as 王明 and 王明树, which leads to more ambiguities. Thirdly, some Chinese person name string has the sense beyond the person name. For example, a common Chinese name, 高峰 has a sense of "*Peak*". Thus, the role of a string as person name or normal word must be determined. Finally, Chinese text is written in continuous character strings without word gap. It leads to the problem that some person names may be segmented into wrong forms.

In the recent years, there have been many researches on person name disambiguation (Fleischman and Hovy 2004; Li et al. 2004; Niu et al. 2004; Bekkerman and McCallum 2005; Chen and Martin 2007; Song et al. 2009). To promote the research in this area, Web People Search (WePS and WePS2) provides a standard evaluation, which focuses on information extraction of personal named-entities in Web data (Artiles et al., 2007; Artiles et al., 2009; Sekine and Artiles, 2009). Generally speaking, both cluster-

based techniques which cluster documents corresponding to one person with similar contexts, global features and document features (Han et al. 2004; Pedersen et al. 2005; Elmacioglu et al. 2007; Pedersen and Anagha 2007; Rao et al. 2007) and information extraction based techniques which recognizes/extracts the description features of one person name (Heyl and Neumann 2007; Chen et al. 2009) are adopted. Considering that these evaluations are only applied to English text, CIPS-SIGHAN 2010 bakeoff proposed the first evaluation campaign on Chinese person name disambiguation. In this evaluation, corresponding to given index person name string, the systems are required to recognize each identical person having the index string as substring and classify the document corresponding to each identical person into a group.

This paper presents the design and implementation of HITSZ_CITYU system in this bakeoff. This system incorporates both recognition/extract technique and clustering technique for person name disambiguation. It consists of two major components. Firstly, by incorporating word segmentation, named entity recognition, and unsupervised clustering, the system recognize the person name string in the document and then classify the documents into subsets corresponding to the person name. Secondly, for the documents having the same person name string, the system identifies the person identify string, other person name, organization name and contextual context words as features. An agglomerative hierarchical clustering algorithm is applied to cluster the documents to each identical person. In this way, the documents corresponding to each identical person are grouped, i.e. the person name ambiguities are removed. The evaluation results show that the HITSZ_CITYU system achieved 0.8399(B-Cubed)/0.8853(P-IP) precisions and 0.9329(B-Cubed)/0.9578(P-IP) recall, respectively. The overall F1 performance 0.8742(B-Cubed)/0.915(P-IP) is ranked 2nd in ten participate teams. These results indicate that the proposed system incorporating both extraction and clustering techniques achieves satisfactory recall and overall performance.

The rest of this report is organized as follows. Section 2 describes and analyzes the task. Section 3 presents the word segmentation and person name recognition and Section 4 presents the person description extraction and document clustering. Section 5 gives and discusses the evaluation results. Finally, Section 6 concludes.

2 Task Description

CIPS-SIGHAN bakeoff on person name disambiguation is a clustering task. Corresponding to 26 person name query string, the systems are required to cluster the documents having the index string into multiple groups, which each group representing a separate entity.

HITSZ_CITYU system divided the whole task into two subtasks:

1. Person name recognition. It includes:

   1.1 Distinguish person name/ non person name in the document. For a given index string 高峰, in Example 1, 高峰 is a person name while in Example 2, 高峰 is a noun meaning "*peak*" rather than a person name.

   Example 1. 谈判专家、北京人民警察学院教授高峰说。(*Gaofeng, the Negotiator and professor of Beijing People's Police College, said*).

   Example 2. 这一数字上升至 11.83%的高峰值 (*This value raise to the peak value of 11.83%*).

   1.2 Recognize the exact person name, especially for three-character to four-character names. For a given index string, 李燕, a person name 李燕 should be identified in Example 3 while 李燕卿 should be identified from Example 4.

   Example 3. 中国一队的李燕是参赛女选手中身材最高的(*Li Yan from Chinese team one is the highest one in the female athletes participating this game*).

   Example 4. 战士李燕卿是个孤儿 (*The soldier Li YanQing is an orphan*)

2. Cluster the documents for each identical person. That is for each person recognized person name, cluster documents into groups while each group representing an individual person. For the non person names instances (such as Example 2), they are clustered into a *discarded* group. Meanwhile, the different person with the same name should be separated. For example, 李燕 in the Example 3 and Example 5 is a athlete and a painter, re-

spectively. These two sentences should be cluster into different groups.

Example 5. 参与主办这次画展的著名画家李燕说(*The famous painter Li Yan , who involved in hosting this exhibition, said that*)

3 Person Name Recognition

As discussed in Section 2, HITSZ_CITYU system firstly recognizes the person names from the text including distinguish the person name/ non-person name word and recognize the different person name having the name index string. In our study, we adopted three developed word segmentation and named entity recognition tools to generate the person name candidates. The three tools are:

1. Language Processing Toolkit from Intelligent Technology & Natural Language Processing Lab (ITNLP) of Harbin Institute of Technology (HIT). http://www.insun.hit.edu.cn/
2. ICTCLAS from Chinese Academy of Sciences. http://ictclas.org/
3. The Language Technology Platform from Information Retrieval Lab of Harbin Institute of Technology. http://ir.hit.edu.cn

We apply the three tools to segment and tag the documents into Chinese words. The recognized person name having the name index string will be labeled as */nr* while the index string is labeled as *discard* if it is no recognized as a person name even not a word. For the sentences having no name index string, we simply vote the word segmentation results by as the output. As for the sentences having name index string, we conduct further analysis on the word segmentation results.

1. For the cases that the matched string is recognized as person name and non-person name by different systems, respectively, we selected the recognized person name as the output. For example, in

   Example 6. 卫生福利及食物局局长**杨永强**赞扬谢婉雯工作尽心尽力 (*Secretary for Health, Welfare and Food, Yang Yongqiang commended the excellent work of Tse Wanwen*).

   the segmentation results by three segmentors are **杨永强/nr |discarded|杨永强/nr**, respectively. We select **杨永强/nr** as the output.
2. For the cases that three systems generate different person names, we further incorporating unsupervised clustering results for determination. Here, an agglomerative hierarchical clustering with high threshold is applied (the details of clustering will be presented in Section 4).

   Example 7. 朱芳勇闯三关 (*Zhufang overcome three barriers*)

   In this example, the word segmentation results are **朱芳/nr, 朱芳勇/nr, 朱芳勇/nr**, respectively. It is shown that there is a segmentation ambiguity here because both 朱芳 and 朱芳勇 are legal Chinese person names. Such kinds of ambiguity cannot be solved by segmentors individually. We further consider the clustering results. Since the Example 7 is clustered with the documents having the segmentation results of 朱芳, two votes (emphasize the clustering confidence) for 朱芳 are assigned. Thus, 朱芳 and 朱芳勇 obtained 3 votes and 2 votes in this case, respectively, and thus 朱芳 is selected as the output.
3. For cases that the different person name forms having the same votes, the longer person name is selected. In the following example,

   Example 8. 上海市教育委员会副主任张民选教授在论坛上表示 (*Prof. Zhang Mingxuan, the deputy director of Shanghai Municipal Education Commission, said at the forum*)

   The segmentation form of 张民 and 张民选 received the same votes, thus, the longer one 张民选 is selected as the output.

In this component, we applied three segmentors (normally using the local features only) with the help of clustering to (using both the local and global features) recognize person name in the text with high accuracy. It is important to ensure the recall performance of the final output. Noted, in order to ensure the high precision of clustering, we set a high similarity threshold here.

4 Person Name Disambiguation

4.1 Person Identity Recognition/Extraction

A person is distinguished by its associated attributes in which its identity description is essential. For example, a person name has the identity of 总统 *president* and 农民 *farmer*, respectively, tends to be two different persons. Therefore, in HITSZ_CITYU system, the person identity is extracted based on lexicon and heuristic rules before person name disambiguation.

We have an entity lexicon consisting of 85 suffixes and 248 prefix descriptor for persons as the initial lexicon. We further expand this lexicon through extracting frequently used entity words from Gigaword. Here, we segmented documents in Gigaword into word sequences. For each identified person name, we collect its neighboring nouns. The associations between the nouns and person name can be estimated by their χ^2 test value. For a candidate entity w_a and person name w_b, (here, w_b is corresponding to person name class with the label */nr*), the following 2-by-2 table shown the dependence of their occurrence.

Table 1 The co-occurrence of two words

	$x = w_a$	$x \neq w_a$
$y = w_b$	C_{11}	C_{12}
$y \neq w_b$	C_{21}	C_{22}

For w_a and w_b, χ^2 test (chi-square test) estimates the differences between observed and expected values as follows:

$$\chi^2 = \frac{N \cdot (C_{11}C_{22} - C_{12}C_{21})^2}{(C_{11} + C_{22}) + (C_{11} + C_{21}) + (C_{12} + C_{22}) + (C_{21} + C_{22})} \quad (1)$$

where, N is the total number of words in the corpus. The nouns having the χ^2 value greater than a threshold are extracted as entity descriptors.

In person entity extraction subtask, for each sentence has the recognized person name, the system matches its neighboring nouns (-2 to +2 words surrounding the person name) with the entries in entity descriptor lexicon. The matched entity descriptors are extracted.

In this part, several heuristic rules are applied to handle some non-neighboring cases. Two example rules with cases are given below.

Example Rule 1. The prefix entity descriptor will be assigned to parallel person names with the split mark of "/" , "、 "and "和","与"(*and*).

中国选手龚跃春/王辉 (*Chinese players Gong Yuechun/Wang Hui*)–>

选手 *player*-龚跃春 *Gong Yuechun*

选手 *player*-王辉 *Wang Hui*

Example Rule 2. The entity descriptor will be assigned to each person in the structure of parallel person name following "等(*etc.*)" and then a entity word.

刘炳森、陈大章、李燕、金鸿钧等书画家挥毫泼墨 (*The painter, Liu Bingsen, Chen Dazhang, Li Yan, Jin Hongjun, etc., paint a.*.) ->

刘炳森 *Liu Bingsen* - 书画家 *painter*

陈大章 *Chen Dazhang* - 书画家 *painter*

李燕 *Li Yan* - 书画家 *painter*

金鸿钧 *Jin Hongjun* - 书画家 *painter*

Furthermore, the HITSZ_CITYU system applies several rules to identify a special kind of person entity, i.e. the reporter or author using structure information. For example, in the beginning or the end of a document, there is a person name in a bracket means this person and this name appear in the document for only once; such person name is regarded as the reporter or author. (记者金林鹏、石涛) –>金林鹏 *Jin Linpeng* - 记者 *reporter*

(金林鹏　李霁) –>金林鹏 *Jin Linpeng* - 记者 *reporter*

4.2 Clustering-based Person Name Disambiguation

For the document set corresponding to each given index person name, we firstly split the document set into: (1) Discarded subset, (2) Subset with different recognized person name. The subsets are further split into (2-1) the person is the author/reporter and (2-2) the person is not the author/reporter. The clustering techniques are then applied to group documents in each (2-2) subset into several clusters which each cluster is corresponding to each identical person.

In the Chinese Person Name Disambiguation task, the number of clusters contained in a subset is not pre-available. Thus, the clustering method which fixes the number of clusters, such as *k-nearest neighbor* (*k-NN*) is not applicable. Considering that Agglomerative Hierarchical Clustering (AHC) algorithm doesn't require the fixed number of cluster and it performs well in docu-

ment categorization (Jain and Dubes 1988), it is adopted in HITSZ_CITYU system.

Preprocessing and Document Representation

Before representing documents, a series of procedures are adopted to preprocess these documents including stop word removal. Next, we select feature words for document clustering. Generally, paragraphs containing the target person name usually contain more person-related information, such as descriptor, occupation, affiliation, and partners. Therefore, larger weights should be assigned to these words. Furthermore, we further consider the appearance position of the features. Intuitively, local feature words with small distance are more important than the global features words with longer distance.

We implemented some experiments on the training data to verify our point. Table 2 and Table 3 show the clustering performance achieved using different combination of global features and local features as well as different similarity thresholds.

Table 2. Performance achieved on training set with different weights (similarity threshold 0.1)

Feature words	Precision	Recall	F-1
Paragraph	0.820	0.889	0.849
All	0.791	0.880	0.826
All+ Paragraph×1	0.791	0.904	0.839
All+ Paragraph×2	0.802	0.908	0.848
All+ Paragraph×3	0.824	0.909	0.860
All+ Paragraph×4	0.831	0.911	0.865
All+ Paragraph×5	0.839	0.910	0.869
All+ Paragraph×6	0.833	0.905	0.864
All+ Paragraph×7	0.838	0.904	0.867

Table 3. Performance achieved on training set with different weights (similarity threshold 0.15)

Feature words	Precision	Recall	F-1
Paragraph	0. 901	0.873	0.883
All	0.859	0.867	0.859
All+ Paragraph×1	0.875	0.887	0.877
All+ Paragraph×2	0.885	0.890	0.884
All+ Paragraph×3	0.889	0.887	0.885
All+ Paragraph×4	0.896	0.887	0.880
All+ Paragraph×5	0.906	0.882	0.891
All+ Paragraph×6	0.905	0.884	0.891
All+ Paragraph×7	0.910	0.882	0.893

In this two tables, "Paragraph" means that we only select words containing in paragraph which contains the person index name as feature words (which are the local features), and "All" means that we select all words but stop words in a document as feature words. "*All+ Paragraph*×*k*" means feature words consist of two parts, one part is obtained from "All", the other is gained from "Paragraph", at the same time, we assign the feature weights to the two parts, respectively. The feature weight coefficient of "All" is $1/(k+1)$, while the feature weight coefficient of "*All+ Paragraph*×*k*" is $k/(k+1)$.

It is shown that, the system perform best using appropriate feature weight coefficient distribution. Therefore, we select all words in the document (besides stop words) as global feature words and the words in paragraph having the index person name as local feature words. We then assign the corresponding empirical feature weight coefficient to the global/local features, respectively. A document is now represented as a vector of feature words as follows:

$$V(d) \to ((t_1, w_1(d)); (t_2, w_2(d)); \cdots (t_n, w_n(d))) \quad (2)$$

where, d is a document, t_i is a feature word, $w_i(d)$ is the feature weight of t_i in the document d. In this paper, we adopt a widely used weighting scheme, named Term Frequency with Inverse Document Frequency (TF-IDF). In addition, for each document, we need to normalize weights of features because documents have different lengths. The weight of word t_i in document d is shown as:

$$w_i(d) = \frac{tf_i(d) * \log(\frac{N}{df_i} + 0.05)}{\sqrt{\sum_{i=1}^{n} (tf_i(d) * \log(\frac{N}{df_i} + 0.05))^2}} \quad (3)$$

where $tf_i(d)$ means how many times word t_i occurs in the document d, df_i means how many documents contains word t_i, and N is the number of documents in the corpus.

Similarity Estimation

We use the cosine distance as similarity calculation function. After the normalization of weights of each document, the similarity between document d_1 and document d_2 is computed as:

$$sim(d_1, d_2) = \sum_{t_i \in d_1 \cap d_2} w_i(d_1) * w_i(d_2) \quad (4)$$

where t_i is the term which appears in document d_1 and document d_2 simultaneously, $w_i(d_1)$ and $w_i(d_1)$ are the weights of t_i in document d_1 and document d_2 respectively. If t_i does not appear in a document, the corresponding weight in the document is zero.

Agglomerative Hierarchical Clustering (AHC)

AHC is a bottom-up hierarchical clustering method. The framework of AHC is described as follows:

Assign each document to a single cluster.
Calculate all pair-wise similarities between clusters.
Construct a distance matrix using the similarity values.
Look for the pair of clusters with the largest similarity.
Remove the pair from the matrix and merge them.
Evaluate all similarities from this new cluster to all other clusters, and update the matrix.
Repeat until the largest similarity in the matrix is smaller than some similarity criteria.

There are three methods to estimate the similarity between two different clusters during the cluster mergence: single link method, average link method and complete link method (Nallapati et al. 2004). The three methods define the similarity between two clusters c_1 and c_2 as follows:

Single link method: The similarity is the largest of all similarities of all pairs of documents between clusters c_1 and c_2 and defined as:

$$sim(c_1, c_2) = \max_{d_i \in c_1, d_j \in c_2} sim(d_i, d_j) \quad (5)$$

Average link method: The similarity is the average of the similarities of all pairs of documents between clusters c_1 and c_2 and defined as:

$$sim(c_1, c_2) = \frac{\sum_{d_i \in c_1} \sum_{d_j \in c_2} sim(d_i, d_j)}{|c_1| * |c_2|} \quad (6)$$

Complete link method: The similarity is the smallest of all similarities of all pairs of documents between clusters c_1 and c_2 and defined as:

$$sim(c_1, c_2) = \min_{d_i \in c_1, d_j \in c_2} sim(d_i, d_j) \quad (7)$$

where, d_i and d_j are the documents belongs to clusters c_1 and c_2, respectively.

We evaluated the AHC algorithm with the above three link methods. The achieved performance are given in Table 4. It is shown that the system performs best with the complete link method. Therefore, the complete link method is selected for the bakeoff testing.

Table 4. Performance achieved on training set with different link method

Similarity threshold	Link method	Precision	Recall	F1
0.1	Single link	0.048	1.000	0.089
0.1	Average link	0.839	0.910	0.869
0.1	Complete link	0.867	0.888	0.874
0.15	Single link	0.048	1.000	0.089
0.15	Average link	0.906	0.882	0.891
0.15	Complete link	0.923	0.868	0.891

5 Evaluations

The task organizer provides two set of evaluation criteria. They are purity-based score (usually used in IR), B-cubed score (used in WePS-2), respectively. The details of the evaluation criteria are given in the task overview.

The performance achieved by the top-3 systems are shown in Table 5.

Table 5. Performance of Top-3 Systems

	B-Cubed			P-IP		
System	Precision	Recall	F1	Precision	Recall	F1
NEU	**0.957**	0.883	**0.914**	**0.969**	0.925	**0.945**
HITSZ	0.839	**0.932**	0.874	0.885	**0.958**	0.915
DLUT	0.826	0.913	0.863	0.879	0.942	0.907

The evaluation results show that the HITSZ_CITYU system achieved overall F1 performance of 0.8742(B-Cubed)/ 0.915(P-IP), respectively.

It is also shown that HITSZ_CITYU achieves the highest the recall performance. It shows that the proposed system is good at split the document to different identical persons. Meanwhile, this system should improve the capacity on merge small clusters to enhance the precision and overall performance.

6 Conclusions

The presented HITSZ_CITYU system applies multi-segmentor and unsupervised clustering to achieve good accuracy on person name string recognition. The system then incorporates entity descriptor extraction, feature word extraction and agglomerative hierarchical clustering method for person name disambiguation. The achieved encouraging performance shown the high performance word segmentation/name recognition and extraction-based technique are helpful to improve the cluster-based person name disambiguation.

References

Andrea Heyl and Günter Neumann. DFKI2: An Information Extraction based Approach to People Disambiguation. Proceedings of ACL SEMEVAL 2007, 137-140, 2007.

Artiles, Javier, Julio Gonzalo and Satoshi Sekine, The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task, Proceedings of Semeval 2007, Association for Computational Linguistics, 2007.

Artiles, Javier, Julio Gonzalo and Satoshi Sekine. "WePS 2 Evaluation Campaign: Overview of the Web People Search Clustering Task, In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009

Bekkerman, Ron and McCallum, Andrew, Disambiguating Web Appearances of People in a Social Network, Proceedings of WWW2005, pp.463-470, 2005

Ergin Elmacioglu, Yee Fan Tan, Su Yan, Min-Yen Kan, and Dongwon Lee. PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features. Proceedings of ACL SEMEVAL 2007, 268-271, 2007.

Fei Song, Robin Cohen, Song Lin, Web People Search Based on Locality and Relative Similarity Measures, Proceedings of WWW 2009

Fleischman M. B. and Hovy E., Multi-document Person Name Resolution, Proceedings of ACL-42, Reference Resolution Workshop, 2004

Hui Han , Lee Giles , Hongyuan Zha , Cheng Li , Kostas Tsioutsiouliklis, Two Supervised Learning Approaches for Name Disambiguation in Author Citations, Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries, 2004

Jain, A. K. and Dubes, R.C. Algorithms for Clustering Data, Prentice Hall, Upper Saddle River, N.J., 1988

Nallapati, R., Feng, A., Peng, F., Allan, J., Event Threading within News Topics, Proceedings of-CIKM 2004, pp. 446–453, 2004

Niu, Cheng, Wei Li, and Rohini K. Srihari,Weakly Supervised Learning for Cross-document Person Name Disambiguation Supported by Information Extraction, Proceedings of ACL 2004

Pedersen, Ted, Amruta Purandare, and Anagha Kulkarni, Name Discrimination by Clustering Similar Contexts, Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico, 2005

Pedersen, Ted and Anagha Kulkarni, Unsupervised Discrimination of Person Names in Web Contexts, Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, 2007.

Rao, Delip, Nikesh Garera and David Yarowsky, JHU1: An Unsupervised Approach to Person Name Disambiguation using Web Snippets, In Proceedings of ACL Semeval 2007

Sekine, Satoshi and Javier Artiles. WePS 2 Evaluation Campaign: overview of the Web People Search Attribute Extraction Task, Proceedings of 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009

Xin Li, Paul Morie, and Dan Roth, Robust Reading: Identification and Tracing of Ambiguous Names, Proceedings of NAACL,pp. 17-24, 2004.

Ying Chen, Sophia Yat Mei Lee, Chu-Ren Huang, PolyUHK: A Robust Information Extraction System for Web Personal Names, Proceedings of WWW 2009

Ying Chen and Martin J.H. CU-COMSEM: Exploring Rich Features for Unsupervised Web Personal Name Disambiguation, Proceedings of ACL Semeval 2007