The Chinese Persons Name Disambiguation Evaluation: Exploration of Personal Name Disambiguation in Chinese News

Ying Chen*, Peng Jin†, Wenjie Li‡, Chu-Ren Huang‡

* China Agricultural University
chenying3176@gmail.com

†Leshan Teachers' College
jandp@pku.edu.cn

‡The Hong Kong Polytechnic University
cswjli@comp.polyu.edu.hk
churenhuang@gmail.com

Abstract

Personal name disambiguation becomes hot as it provides a way to incorporate semantic understanding into information retrieval. In this campaign, we explore Chinese personal name disambiguation in news. In order to examine how well disambiguation technologies work, we concentrate on news articles, which is well-formatted and whose genre is well-studied. We then design a diagnosis test to explore the impact of Chinese word segmentation to personal name disambiguation.

1 Introduction

Incorporating semantic understanding technologies from the field of NLP becomes one of further directions for information retrieval. Among them, named entity disambiguation, which intends to use state-of-the-art named entity processing to enhance a search engine, is a hot research issue. Because of the popularity of personal names in queries, more efforts are put on personal name disambiguation. The personal name disambiguation used both in Web Personal Search (WePS[1]) and our campaign is defined as follow. Given documents containing a personal name in interest, the task is to cluster them according to which entity the name in a document refers to.

WePS, which explores English personal name disambiguation, has been held twice (Artiles et al., 2007, 2009). Compared to the one in English, personal name disambiguation in Chinese has special issues, such as Chinese text processing and Chinese personal naming system. Therefore, we hold Chinese personal name disambiguation (CPND) to explore those problems. In this campaign, we mainly examine the relationships between Chinese word segmentation and Chinese personal name disambiguation.

Moreover, from our experiences in WePS (Chen et al., 2007, 2009), we notice that webpages are so noisy that text pre-processing that extracts useful text for disambiguation needs much effort. In fact, text pre-processing for webpages is rather complicated, such as deleting of HTML tags, the detection of JavaScript codes and so on. Therefore, the final system performance in the WePS campaign sometimes does not reflect the disambiguation power of the system, and instead it shows the comprehensive result of text pre-processing as well as disambiguation. In order to focus on personal name disambiguation, we choose news documents in CPND.

The paper is organized as follows. Section 2 describes our formal test including datasets and evaluation. Section 3 introduces the diagnosis test, which explores the impact of Chinese word segmentation to personal name disambiguation. Section 4 describes our campaign, and Section 5 presents the results of the participating systems. Finally, Section 6 concludes our main findings in this campaign.

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

2 The Formal Test

2.1 Datasets

To avoid the difficulty to clean a webpage, we choose news articles in this campaign. Given a full name in Chinese, we search the character-based personal name string in all documents of Chinese Gigaword Corpus, a large Chinese news collection. If a document contains the name, it is belonged to the dataset of this name. To ensure the popularity of a personal name, we keep only a personal name whose corresponding dataset comprises more than 100 documents. In addition, if there are more than 300 documents in that dataset, we randomly select 300 articles to annotate. Finally, there are totally 58 personal names and 12,534 news articles used in our data, where 32 names are in the development data and 26 names are in the test data, as shown Appendix Table 4 and 5 separately.

From Table 4 and 5, we can find that the ambiguity (the document number per cluster) distribution is much different between the development data and the test data. In fact, the ambiguity varies with a personal name in interest, such as the popularity of the name in the given corpus, the celebrity degree of the name, and so on.

2.2 Evaluation

In WePS, Artiles et al. (2009) made an intensive study of clustering evaluation metrics, and found that B-Cubed metric is an appropriate evaluation approach. Moreover, in order to handle overlapping clusters (i.e. a personal name in a document refers to more than one person entity in reality), we extend B-Cubed metric as Table 1, where S = {S_1, S_2, …} is a system clustering and R = {R_1, R_2, …} is a gold-standard clustering. The final performance of a system clustering for a personal name is the F score (α = 0.5), and the final performance of a system is the Mac F score, the average of the F scores of all personal names.

Moreover, Artiles et al. (2009) also discuss three cheat systems: one-in-one, all-in-one, and the hybrid cheat system. One-in-one assigns each document into a cluster, and in contrast, all-in-one put all documents into one cluster. The hybrid cheat system just incorporates all clusters both in one-in-one and all-in-one clustering. Although the hybrid cheat system can achieve fairly good performance, it is not useful for real applications. In the formal test, these three systems serve as the baseline.

	Formula
Precision	$\dfrac{\sum_{S_i \in S} \sum_{d \in S_i} \max_{R_j \in R; d \in R_j} \dfrac{\mid S_i \cap R_j \mid}{\mid S_i \mid}}{\sum_{S_i \in S} \mid S_i \mid}$
Recall	$\dfrac{\sum_{R_i \in R} \sum_{d \in R_i} \max_{S_j \in S; d \in S_j} \dfrac{\mid R_i \cap S_j \mid}{\mid R_i \mid}}{\sum_{R_i \in R} \mid R_i \mid}$

Table 1: the formula of the modified B-cubed metrics

3 The Diagnosis Test

Because of no word delimiter, Chinese text processing often needs to do Chinese word segmentation first. In order to explore the relationship between personal name disambiguation and word segmentation, we provide a diagnosis data which attempts to examine the impact of word segmentation to disambiguation.

Firstly, for each personal name, its corresponding dataset will be manually divided into three groups as follows. The disambiguation system then runs for each group of documents. The three clustering outputs are merged into the final clustering for that personal name.

(1) Exactly matching: news articles containing personal names that exactly match the query personal name.

(2) Partially matching: news articles containing personal names that are super-strings of the query personal name. For instance, an article that has a person named with “高军田” (*Gao Jun* Tian) is retrieved for the query personal name “高军” (Gao Jun).

(3) Discarded: news articles containing character sequences that match the query personal name string and however in fact are not a personal name. For instance, an article that has the string “最高军事法

院” (Zui *Gao Jun* Shi Fa Yuan: supreme military court) is also retrieved for the personal name “高军” (Gao Jun).

This diagnosis test is designed to simulate the realistic scenario where Chinese word segmentation works before personal name disambiguation. If a Chinese word segmenter works perfectly, a word-based matching can be used to retrieve the documents containing a personal name, and articles in Groups (2) and (3) should not be returned. The personal name disambiguation task that is limited to the documents in Group (1) should be simpler.

Moreover, in this diagnosis test, we propose a baseline based on the gold-standard word segmentation as follows, namely the word-segment system.

1) All articles in the “exactly matching” group are merged into a cluster, and all articles in the “discarded” group are merged into a cluster.
2) In the “partially matching” group, entities exactly sharing the same personal name are merged into a cluster. For example, all articles containing “高军田” (Gao Jun Tian) are merged into a cluster, and all articles containing 高军华” (Gao Jun Hua) are merged into another cluster.

4 Campaign Design

4.1 The Participants

The task of Chinese personal name disambiguation in news has attracted the participation of 10 teams. As a team can submit at most 2 results, there are 17 submissions from the 10 teams in the formal test, and there are 11 submissions from 7 teams in the diagnosis.

4.2 System descriptions

Regarding system architecture, all systems are based on clustering, and most of them comprise two components: feature extraction and clustering. However, NEU-1 and HITSZ_CITYU develop a different clustering, which in fact is a cascaded clustering. Taking the advantage of the properties of a news article, both systems first divide the dataset for a personal name into two groups according to whether the person in question is a reporter of the news. They then choose a different strategy to make further clustering for each group.

In terms of feature extraction, we find that all systems except SoochowHY use word segmentation as pre-processing. Moreover, most systems choose named entity detection to enhance their feature extraction. In addition, character-based bigrams are also used in some systems. In Appendix Table 6, we give the summary of word segmentation and named entity detection used in the participating systems.

Regarding clustering algorithms, agglomerative hierarchical clustering is popular in the submissions. Moreover, we find that weight learning is very crucial for similarity matrix, which has a big impact to the final clustering performance. Besides the popular Boolean and TFIDF weighting schemes, SoochowHY and NEU-2 use different weighting learning. NEU-2 manually assigns weights to different kinds of features. SoochowHY develops an algorithm that iteratively learns a weight for a character-based n-gram.

5 Results

We first provide the performances of the formal test, and make some analysis. We then present and discuss the performances of the diagnosis test.

5.1 Results of the Formal test

For the formal test, we show the performances of 11 submissions from 10 teams in Table 2. For each team, we keep only the better result except the NEU team because they use different technologies in their two submissions (NEU_1 and NEU_2).

From Table 2, we first observe that 7 submissions perform better than the hybrid cheat system. In contrast, in Artiles et al. (2009), only 3 teams can beat the hybrid system. From our analysis, this may attribute to the following facts.

1) Personal name disambiguation on Chinese may be easier than the one on English. For example, one of key issues in personal name disambiguation is to capture the occurrences of a query name in text. However, various personal name expressions, such as the use of

	Precision	Recall	Macro F
NEU_1	95.76	88.37	91.47
NEU_2	95.08	88.62	91.15
HITSZ_CITYU	83.99	93.29	87.42
ICL_1	83.68	92.23	86.94
DLUT_1	82.69	91.33	86.36
BUPT_1	80.33	94.52	85.79
XMU	90.55	84.88	85.72
Hybrid cheat system	73.48	100	82.37
HIT_ITNLP_2	91.08	62.75	71.03
BIT	80.2	68.75	68.4
ALL_IN_ONE	52.54	100	61.74
BUPT_pris02	72.39	58.35	57.68
SoochowHY_2	84.51	44.17	51.42
ONE_IN_ONE	94.42	14.41	21.07

Table 2: The B-Cubed performances of the formal test

	Precision	Recall	Macro F
NEU_1	95.6	89.74	92.14
NEU_2	94.53	89.99	91.66
XMU	89.84	89.84	89.08
ICL_1	84.53	93.42	87.96
BUPT_1	80.43	95.41	86.18
Word_segment system	71.11	100	80.92
BUPT_pris01	77.91	75.09	74.25
BIT	94.62	63.32	72.48
SoochowHY	87.22	58.52	61.85

Table 3: The B-Cubed performances of the diagnosis test

middle names in English, cause many problems during recognizing of the occurrences of a personal name in interest.

2) We works on news articles, which have less noisy information compared to webpages used in Artiles et al. (2009). More efforts are put on the exploration directly on disambiguation, not on text pre-processing. Furthermore, most of systems extract features based on some popular NLP techniques, such as Chinese word segmentation, named entity recognition and POS tagger. As those tools usually are developed based on news corpora, they should extract high-quality features for disambiguation in our task.

We then notice that the NEU team achieves the best performance. From their system description, we find that they make some special processing just for this task. For example, they develop a personal name recognition system to detect the occurrences of a query name in a news article, and a cascaded clustering for different kinds of persons.

5.2 Results of the Diagnosis test

We present the performances of 8 submissions for the diagnosis test from 7 teams in Table 3 as the format of Table 2. Meanwhile, we use the word-segment system as the baseline.

Comparing Table 2 and 3, we first find that the word-segment system has a lower performance than the hybrid cheat system although the word-segment system is more useful for real applications. This implies the importance to develop an appropriate evaluation method for clustering. From Table 3, five submissions achieve better performances than the word-segment system.

Given the gold-standard word segmentation on personal names in the diagnosis test, from Table 3, our total impression is that the top systems take less advantages, and the bottom systems take more. This indicates that bottom systems suffer from their low-quality word segmentation and named entity detection. For example, BUPT_pris01 increases ~22% F score (from 52.81% to 74.25%).

6 Conclusions

This campaign follows the work of WePS, and explores Chinese personal name disambiguation on news. We examine two issues: one is for Chinese word segmentation, and the other is noisy information. As Chinese word segmentation usually is a pre-processing for most NLP processing, we investigate the impact of word segmentation to disambiguation. To avoid noisy information for disambiguation, such as HTML tags in webpage used in WePS, we choose news article to work on so that we can capture how good the state-of-the-art disambiguation technique is.

References

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

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*.

Bagga, Amit and Breck Baldwin.1998. Entity-based Cross-document Co-referencing Using the Vector Space Model. In Proceedings of *the 17th International Conference on Computational Linguistics*.

Chen, Ying and James H. Martin. 2007. CU-COMSEM: Exploring Rich Features for Unsupervised Web Personal Name Disambiguation. In Proceedings of *Semeval 2007, Association for Computational Linguistics*.

Chen, Ying, 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*.

Appendix

name	document #	cluster #	document # per cluster
赵伟	155	37	4.19
高明	301	42	7.17
高军	300	5	60
高伟	105	30	3.5
郭伟	156	42	3.71
朱建军	350	15	23.33
杨伟	269	70	3.84
何涛	257	8	32.13
王华	211	109	1.94
马杰	177	36	4.92
罗杰	358	165	2.17
黄海	300	20	15
徐明	140	57	2.46
刘海	300	27	11.11
孙海	296	73	4.05
黄明	135	75	1.8
郭勇	297	14	21.21
唐海	110	24	4.58
孙明	207	68	3.04

何海	131	26	5.04
郭华	145	22	6.59
孙涛	164	15	10.93
张建军	247	20	12.35
杨波	173	34	5.09
张志强	171	21	8.14
梁伟	170	34	5
胡明	195	32	6.09
林海	301	22	13.68
李刚	318	76	4.18
李军	234	117	2
胡刚	134	9	14.89
马强	123	7	17.57
	6930	1352	5.13

Table 4: The training data distribution

name	document #	cluster #	document # per cluster
刘俊	190	96	1.99
郭超	191	5	38.2
罗毅	258	16	16.13
王建民	224	32	7
王晓东	118	29	4.07
赵颖	239	21	11.38
王峰	208	43	4.84
李建民	201	17	11.82
黄志红	317	3	105.67
杨永强	151	6	25.17
何文	188	61	3.08
李学军	200	2	100
李燕	213	69	3.09
刘洪波	182	5	36.4
林鹏	278	11	25.27
周雷	180	4	45
徐金平	286	1	286
李玲	206	38	5.42
孙平	193	16	12.06
吴小军	172	9	19.11
朱芳	174	5	34.8
张民	299	39	7.67
刘丽	233	90	2.59
高峰	300	13	23.08
朱洪	141	25	5.64

王永康	262	13	20.15
	5604	669	8.38

Table5: The test data distribution

	Word segmentation	Named Entity
NEU	Name: Neucsp Source: 1998 People's Daily	Name: in-house
HITSZ_CITYU		
ICL	Name: LTP F score: 96.5% Source: 2nd SIGHAN	Name: LTP
DLUT		
BUPT	Name: in-house F score: 96.5% Source: SIGHAN 2010	
XMU	Name: in-house Source: 1998 People's Daily F score: 97.8%	
HIT_ITNLP	Name: IRLAS Source: 1998 People's Daily F score: 97.4%	Name: IRLAS
BIT	Name: ICTCLAS2010 Precision: ~97% Source: 1998 People's Daily	Name: ICTCLAS2010
BUPT_pris	Name: LTP	Name: LTP
SoochowHY	None	None

Table 6: The summary of word segmentation and named entity detection used in the participants

* LTP(Language Technology Platform)