The Task 2 of CIPS-SIGHAN 2012
Named Entity Recognition and Disambiguation in Chinese Bakeoff

Zhengyan He Houfeng Wang * **Sujian Li**
MOE Key Lab of Computational Linguistics, Peking University
hezhengyan.hit@gmail.com, {wanghf, lisujian}@pku.edu.cn

Abstract

The CIPS-SIGHAN 2012 Chinese Named Entity Recognition and Disambiguation (NERD) bake-off was held in the summer of 2012. Named entity recognition and disambiguation is an important task in natural language processing and knowledge base construction. It aims at detecting entity mentions in raw text, followed by pointing the detected mentions to real world entities. Often, real world entities can be found on online encyclopedia like Wikipedia and Baike. This task focuses on NERD in Chinese Language, and presents some challenges unique to Chinese, namely the confusion of named entity with common words, and lack of capital clues as in English. We manually construct query names and a knowledge base from Baike. Evaluation results show promising future of this field.

1 Overview

Named Entity Recognition and Disambiguation (NERD) is the task of detecting entity mentions from raw text and classifying each mention to its real world entity. NERD is a fundamental problem in Natural Language Processing (NLP), and the first step towards many higher level tasks, such as constructing knowledge bases, populating entities with attributes, social analysis, information extraction and question answering.

NERD in Chinese has posed some unique challenges. First, common words can be used as named entities. For example, 高明(brilliant), a common adjective, is also a person name in China. Therefore, it is challenging to distinguish common words which function as named entities, given that Chinese words have less morphology variations than many other languages. Second, different types of named entities can use the same names. For example, 金山(Gold Hill) can be used as the name of persons, locations and organizations. Finally, it is typical in China that many persons share the same name. For instance, there are many persons having the name 王刚(Wang Gang) in China. To investigate these issues, SIGHAN 2012 establishes a task for Named Entity Recognition and Disambiguation (NERD task).

Similar tasks in English have been studied for several years. Related events include Knowledge Base Population (KBP) track of Text Analysis Conference (TAC) (Ji and Grishman, 2011; Ji et al., 2010), Web People Search (WePS) (Artiles et al., 2007). In WePS, the task is person name clustering, in which there is no knowledge base available. In TAC-KBP, the task is called entity linking, where the knowledge base is constructed with a subset of Wikipedia, and an entity linking system should output the correct entity id in knowledge base or "NIL" if the entity is not present in the knowledge base . It is also closely related to cross-document coreference resolution. Some other names like entity disambiguation (Kataria et al., 2011) and Wikification (Mihalcea and Csomai, 2007) are also used.

In the SIGHAN 2012 NERD task, 8 teams has successfully submitted their results and several approaches have proved to be quite effective and promising.

* corresponding author

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 108–114,
Tianjin, China, 20-21 DEC. 2012

2 Task Definition and Evaluation Metrics

2.1 Task description

The participants are provided with a collection of web documents (the Source) and a Knowledge Base (KB) which contains the targets of disambiguation. One needs to find for each mention the target entity it refers to, according to the context in which it appears.

Table 1 is a sample of the knowledge base. Each one is an XML document, in which there are several candidate entities with the same name, and each entity has a short description. Each ambiguous name has a collection of test text. For each test text, one should determine which real entity the name refers to, if it presents in the knowledge base, output the id in the KB; or if it is a common word, output "Other"; or if it is an entity outside the KB, group them into different clusters, output "Out_n".

2.2 dataset preparation

The query person names are manually selected to reflect both the variation of this name and the confusion with common words. knowledge base is constructed from Baidu Baike entries according the person names. Source texts are selected by 20 student querying the search engine. The students are advised to crawl web document with as many variation of persons for each name as possible, and also with common words. The crawled documents for one query are splitted into folders for each real person in Baike, and reviewed by the advisor.

The query names are chosen to reflect some commonly observed in Chinese person name recognition and disambiguation, such as common words ("张扬" "田野" "高明"), entity type variation ("沈阳" "金山" "黄河").

The entire dataset contains 32 names in Chinese. Table 2 gives an overview of the dataset.

2.3 Evaluation

For each name, there is a collection of test documents for evaluation. Evaluation is carried out on a per document basis. Let T denote the document collection for one name (e.g. "雷雨"), for each query document $t \in T$, the system output may fall into three classes, namely: SL_XX, SOther and SOut_XX, representing in-KB id, a common word,

```
<?xml version='1.0' encoding='UTF-8'?>
<EntityList name="雷雨">
  <Entity id="01">
    <text>通江县第二中学教师，男，大学本科，西华师范大学英语语言文学专业毕业。高
    二英语备课组长。自参工以来一事从事高中英语教学工作，长期从事班主任工作，所任
    班级历届成绩显著。...
    </text>
  </Entity>
  <Entity id="02">
    <text>重庆市黔江区太极乡党委副书记、乡长。主持政府全面工作，主管财政、金融、
    审计、统计、非公有制经济、城乡统筹、乡镇企业、招商引资、烤烟、蚕桑工作。
    </text>
  </Entity>
  <Entity id="03">
    <text>罗源县中房镇下湖村人。1978年8月加入中国共产党。1981年，毕业于上海同济大
    学规划专业。同年起，任福州市城乡设计院规划室主任、工程师，兼任福州市土木建筑
    学会秘书长。...
    </text>
  </Entity>
  <Entity id="04">
    <text>男，汉族，硕士研究生学历，出生于1961年9月，陕西 中共商南县委书记，商州
    人，1980年8月参加革命工作，1982年7月加入中国共产党，现任中共商南县委书记。曾
    任任共青团商洛地委副书记；洛南县政府副县长；任中共商南县委副书记；中共山阳县
    委常委、县政府常务副县长，等。
    </text>
  </Entity>
  <Entity id="05">
    <text>四川省蒲江县教育局党组书记、局长。主持县教育局全面工作。主管教育督导、
    计财、基建和教仪电教等工作。
    </text>
  </Entity>
  <Entity id="06">
    <text>女，1975年8月生，回族，广西南宁人，中共党员，1997年7月广西师范大学汉语
    言专业毕业，2006年获教育硕士学位，中学中级教师，1997年7月进入桂林中学任教语文
    至今。
    </text>
  </Entity>
</EntityList>
```

Table 1: Sample of Knowledge Base. Each entry contains a short description of the real world entity.

Name	in-KB					not-in-KB					Other
	#text	#cluster	max	min	avg	#text	#cluster	max	min	avg	
丛林	81	5	20	7	16.0	14	9	3	1	1.0	24
严明	37	12	13	2	3.0	0	0	0	0	0.0	10
华山	109	9	18	7	12.0	19	4	6	3	4.0	0
华明	55	4	19	6	13.0	10	5	3	1	2.0	0
吉祥	56	8	19	1	7.0	1	1	1	1	1.0	19
张弛	202	27	24	1	7.0	52	12	7	2	4.0	26
张扬	145	19	15	1	7.0	0	0	0	0	0.0	14
方正	115	12	18	1	9.0	12	4	5	1	3.0	4
李晓明	416	33	33	2	12.0	86	15	9	2	5.0	0
杜鹃	155	13	21	2	11.0	12	8	5	1	1.0	12
杨柳	210	15	25	1	14.0	22	5	9	2	4.0	18
江涛	248	28	26	1	8.0	16	6	6	1	2.0	17
汪洋	181	12	37	1	15.0	21	4	8	1	5.0	21
田野	258	34	21	1	7.0	11	2	8	3	5.0	20
白云	244	19	28	2	12.0	16	2	9	7	8.0	18
白雪	116	9	19	5	12.0	0	0	0	0	0.0	17
秦岭	78	12	15	1	6.0	22	2	16	6	11.0	0
约翰逊	254	15	20	3	16.0	74	18	11	2	4.0	12
胡琴	43	3	22	7	14.0	7	3	3	2	2.0	24
金山	115	8	17	9	14.0	5	1	5	5	5.0	5
雷雨	56	6	17	3	9.0	7	1	7	7	7.0	23
马啸	57	6	18	2	9.0	9	2	6	3	4.0	3
高山	126	19	19	1	6.0	4	1	4	4	4.0	20
高峰	200	37	19	1	5.0	3	1	3	3	3.0	24
高明	195	22	20	1	8.0	16	3	11	1	5.0	23
高超	88	13	19	2	6.0	13	7	3	1	1.0	15
高雄	78	4	29	10	19.0	6	2	4	2	3.0	0
黄梅	150	13	22	3	11.0	3	2	2	1	1.0	19
黄河	156	14	26	1	11.0	22	4	8	4	5.0	0
黄海	108	19	15	1	5.0	20	3	8	5	6.0	0
黄莺	80	9	16	4	8.0	15	4	5	2	3.0	24
黄龙	129	14	21	1	9.0	23	4	7	3	5.0	9

Table 2: Statistics of dataset. Each column in in-KB and not-in-KB means number of texts in total, number of entities in total, max/min/average number of texts containing the name. The last column is number of texts classified as "Other" in gold standard.

or a out-of-KB cluster id respectively; the gold label is L_XX, Other and Out_XX. We compute the precision and recall for this query as follows:

1. if t in T is predicted as SL_XX, we use the following formulae.

$$Pre(t) = \frac{|SL\_XX \cap L\_XX|}{|SL\_XX|} \quad (1)$$

$$Rec(t) = \frac{|SL\_XX \cap L\_XX|}{|L\_XX|} \quad (2)$$

2. if t in T is predicted as SOther, we use the following formulae.

$$Pre(t) = \frac{|SOther \cap Other|}{|SOther|} \quad (3)$$

$$Rec(t) = \frac{|SOther \cap Other|}{|Other|} \quad (4)$$

3. if t in T is predicted as SOut_XX, we use the following formulae.

$$Pre(t) = \frac{|SOut\_XX(t) \cap Out\_YY(t)|}{|SOut\_XX|} \quad (5)$$

$$Rec(t) = \frac{|SOut\_XX(t) \cap Out\_YY(t)|}{|Out\_YY|} \quad (6)$$

4. Accorting to all the instance documents of 雷雨, the overall precision and recall are calculated as follows.

$$Pre(n) = \frac{\sum_{t \in T} Pre(t)}{|T|} \quad (7)$$

$$Rec(n) = \frac{\sum_{t \in T} Rec(t)}{|T|} \quad (8)$$

5. The overall precision and recall for all test names are calculated as follows (the set of all the test names are notated as N, each name is represented as n in N)

$$Pre = \frac{\sum_{n} Pre(n)}{|N|} \quad (9)$$

$$Rec = \frac{\sum_{n} Rec(n)}{|N|} \quad (10)$$

$$F = \frac{2 \times Pre \times Rec}{Pre + Rec} \quad (11)$$

Organization	Contact
NLP group at the University of Macau(I)	Longyue Wang
NLP group at the University of Macau(II)	Hao Zong
Shenzhen Graduate School, Harbin Institute of Technology & Hong Kong Polytechnic University	Jian Xu
Kunming University of Science and Technology	Zhengtao Yu
Institute of Automation, Chinese Academy of Sciences	Tao Zhang
Beijing University of Posts and Telecommunications	Caixia Yuan
Zhengzhou University	HongyingZan
Institute of Software, Chinese Academy of Sciences	Le Sun

Table 3: List of participants

3 Participants of this task

Table 3 lists the 8 teams of the bake-off task.

4 Results, System Comparison and Discussion

4.1 Basic steps of recognition and disambiguation

There are several common components shared by many teams, which is determined by the task requirements:

- preprocessing: the KB and Source text are segmented into Chinese words, and other processing like POS-tagging and named entity recognition are alternatively used;
- information extraction: keywords, entities and relevant attributes are extracted, to construct a vector representation of KB and Source text;
- similarity calculation: the similarity is computed with feature vector, and entities in KB is generated by the rank score. Most teams use simply the unsupervised method to rank candidates, and some teams use semantic resources like Tongyici Cilin (Tian et al., 2012) or the Web for a better scoring;

- "NIL" entity clustering: maximum similarity score below a threshold is a good sign of determining if the entity is in the KB. Hierarchical clustering method is used by many teams to group NIL entities (Peng et al., 2012; Zhang et al., 2012).

- a separate common word detection step is used after the first entity recognition step, or after the knowledge base linking phase.

There are several features which proves useful for accurate disambiguation. The features are listed as follows:

- keywords: one team report extracting discriminative keywords from the KB to represent the target entities, besides using bag-of-word feature vector, and the performance is good (Zong et al., 2012).

- entity of different types: person, organization, location, and other types are used by many teams (Qing-hu et al., 2012; Peng et al., 2012; Zong et al., 2012; Wang et al., 2012). One team reports cooccuring persons more discriminative than other types (Zong et al., 2012). This is reasonable since a person is largely influenced by its social relations.

- entity attributes: several teams (Tian et al., 2012; Wang et al., 2012; Wei et al., 2012) extract attribute of many types, such as title, occupation, gender, nationality, graduate school, education background, publication, etc. Whether the performance is good is largely determined by the extraction technique.

- representation of pseudo-entities (i.e. "Other" and "Out_n"): one team benefits from a explicit representation of common words and out-of-KB entities (Peng et al., 2012), rather than using same set of feature for classification and clustering. They leverage the Web to discover keywords frequently occurring with common names. They further make the assumption that if all the entities in test document do not appear in the entries of KB, then it is likely to be an out-of-KB entity.

Feature weighting tuning: with those diverse kinds of representative features, the NERD system has to determine which feature is more important. One team uses supervised method to tune the weight of different features (Tian et al., 2012), while another team uses the information gain criterion (Wei et al., 2012).

Besides a good representation of both source text and knowledge base entities, there are other aspects that may benefit a NERD system. One team use model combination method: there are several rank score and each with different feature input; a classification model finally determine the relative importance of each scoring (Liu et al., 2012). Training set can be used to decide the threshold in NIL linking and tune the weight of different features and models. One team also uses the extended version of KB from Baidu Baike to enrich the feature set (Liu et al., 2012), and constructs a one-to-one mapping from Baike to KB, because most of the entities is constructed from Baike.

4.2 Analysis of difficult queries

Table 4 shows detailed top/median precision/recall/f-score across all teams, for each query name. The result shows that the performance is good for most of the queries, except for a few, like "田野" "黄河" "黄莺" "黄龙". As we did not have the named entity recognition result, we detect it is due to their so common usage in Chinese Language as a common word. It is even harder for the detection system to consider it as a named entity without strong clues.

Table 5 shows detailed median score for in-KB, NIL clustering, and common word detection results. We can see that the precision and recall of in-KB entities are generally much higher than the NIL clustering. This is reasonable because the entities in KB are almost famous people and rich in attributes and cooccurence entities, as most systems use these attributes as strong indicator of specific person.

Moreover, there is general trend that the recall of NIL clustering is higher than precision. That is to say most of the systems tend to put entities into separate clusters. The reason may be that most NIL entities are so rarely observed and have fewer clues like social relations. They are in most situations dissimilar to each other, if the system uses attribute or

name	precision	recall	f-score
丛林	0.867/0.806	0.916/0.783	0.883/0.778
严明	0.972/0.798	0.885/0.724	0.920/0.777
华山	0.809/0.722	0.863/0.723	0.792/0.697
华明	0.969/0.837	0.905/0.866	0.936/0.822
吉祥	0.934/0.833	0.955/0.882	0.938/0.842
张弛	0.750/0.615	0.905/0.830	0.820/0.692
张扬	0.907/0.786	0.915/0.824	0.904/0.807
方正	0.860/0.792	0.926/0.797	0.885/0.738
李晓明	0.859/0.618	0.871/0.720	0.812/0.674
杜鹃	0.870/0.749	0.852/0.793	0.853/0.759
杨柳	0.868/0.785	0.890/0.808	0.855/0.797
江涛	0.836/0.661	0.825/0.778	0.830/0.709
汪洋	0.866/0.675	0.837/0.736	0.847/0.684
田野	0.734/0.649	0.791/0.718	0.761/0.683
白云	0.813/0.660	0.867/0.697	0.819/0.694
白雪	0.925/0.839	0.929/0.846	0.927/0.839
秦岭	0.817/0.680	0.861/0.715	0.837/0.699
约翰逊	0.734/0.621	0.890/0.719	0.804/0.685
胡琴	0.973/0.890	1.000/0.843	0.978/0.850
金山	0.937/0.777	0.925/0.809	0.931/0.767
雷雨	0.942/0.796	0.898/0.766	0.847/0.802
马啸	0.930/0.868	0.911/0.826	0.893/0.843
高山	0.880/0.763	0.874/0.804	0.867/0.796
高峰	0.916/0.746	0.848/0.755	0.880/0.759
高明	0.861/0.709	0.899/0.748	0.871/0.721
高超	0.806/0.672	0.894/0.769	0.822/0.703
高雄	0.917/0.765	0.966/0.732	0.843/0.722
黄梅	0.822/0.803	0.857/0.815	0.831/0.786
黄河	0.729/0.667	0.875/0.727	0.740/0.690
黄海	0.891/0.690	0.929/0.757	0.892/0.738
黄莺	0.783/0.660	0.922/0.760	0.781/0.665
黄龙	0.528/0.340	0.681/0.477	0.447/0.411
total	0.795/0.702	0.856/0.732	0.802/0.721

Table 4: analysis of queries. Each cell gives the maximum/median score over all teams.

cooccuring entities, simply because the features of these types have a small opportunity to match.

Finally, the "Other" class performance differs a lot across different queries. We deduce this is caused by the difficulty level of the query document. As this part is closely related to the segmentation and entity recognition processing step, it is hard to tell which aspects are more important, the recognition or segmentation.

It is interesting to see that with so many difficulty discussed, there are general clues which indicate a good performance of an NERD system. Most systems use fine-grained keywords, attributes, and cooccurence entities, which gives competitive performance. One team exceeds over 80% total F-score, and 3 teams at around 75%. We can expect better performance with better recognition tools and even large collections of Source and KB information.

5 Conclusion

The Chinese named entity recognition and disambiguation task for CIPS-SIGHAN 2012 has raised the problem in Chinese NERD. Besides the basic difficulty of detection, classification, and NIL clustering , there are other difficulties like common words detection, disambiguation across entity types. 8 teams have submitted their results, and address the difficulties in different ways. Most teams use simple unsupervised scoring metrics, with careful design of feature representation. Some of the techniques prove effective and the result is promising.

Acknowledgment

This work was partially supported by National High Technology Research and Development Program of China (863 Program) (No. 2012AA011101), National Natural Science Foundation of China (No.91024009, No.60973053), the Specialized Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20090001110047)

References

J. Artiles, J. Gonzalo, and S. Sekine. 2007. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. *Proceedings of Semeval*, pages 64–69.

Name	in-KB p/r	Out_n p/r	Other p/r
丛林	0.85/0.77/5	0.71/0.83/9	0.93/0.74/24
严明	0.91/0.82/7	1.00/0.00/0	0.76/0.79/6
华山	0.77/0.76/9	0.59/0.87/4	0.00/0.00/0
华明	0.95/0.89/4	0.71/0.85/5	0.00/0.00/0
吉祥	0.81/0.92/8	1.00/1.00/1	1.00/0.73/19
张弛	0.63/0.82/27	0.69/0.77/12	0.83/0.69/26
张扬	0.77/0.83/19	0.79/0.00/0	0.89/0.65/14
方正	0.81/0.82/12	0.71/0.66/4	0.38/0.77/4
李晓明	0.69/0.74/32	0.50/0.66/15	0.00/0.00/0
杜鹃	0.80/0.79/13	0.67/0.83/8	0.88/0.70/12
杨柳	0.82/0.83/15	0.68/0.68/5	0.65/0.65/18
江涛	0.70/0.78/27	0.71/0.83/6	0.21/0.76/17
汪洋	0.69/0.75/12	0.46/0.69/4	0.69/0.59/21
田野	0.66/0.75/32	0.73/0.80/2	0.66/0.54/20
白云	0.77/0.71/19	0.51/0.71/2	0.75/0.59/18
白雪	0.86/0.90/9	0.79/0.00/0	0.89/0.68/17
秦岭	0.81/0.83/10	0.89/0.77/2	0.00/0.00/0
约翰逊	0.72/0.79/15	0.45/0.66/18	0.03/0.62/12
胡琴	0.86/0.97/3	0.69/0.79/3	0.95/0.73/24
金山	0.92/0.83/8	0.53/0.80/1	0.50/0.70/5
雷雨	0.84/0.76/6	0.85/0.69/1	0.93/0.78/23
马啸	0.89/0.85/6	0.78/0.73/2	0.53/0.56/3
高山	0.79/0.85/17	0.85/0.81/1	0.73/0.66/20
高峰	0.78/0.79/31	0.87/0.71/1	0.69/0.64/22
高明	0.81/0.80/18	0.70/0.74/3	0.70/0.65/19
高超	0.69/0.79/12	0.83/0.79/7	0.74/0.75/14
高雄	0.89/0.75/4	0.77/0.72/2	0.00/0.00/0
黄梅	0.82/0.82/13	0.61/0.96/2	0.72/0.63/19
黄河	0.77/0.81/13	0.55/0.88/4	0.00/0.00/0
黄海	0.80/0.80/18	0.55/0.82/3	0.00/0.00/0
黄莺	0.75/0.78/9	0.55/0.74/4	0.59/0.62/24
黄龙	0.34/0.44/15	0.47/0.52/4	0.52/0.65/9

Table 5: Statistics of in-KB, out-of-KB, other class performance; the score is median of precision, recall; and number of types of entity for in-KB and out-of-KB, number of Other documents in gold standard.

H. Ji and R. Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1148–1158.

H. Ji, R. Grishman, H.T. Dang, K. Griffitt, and J. Ellis. 2010. Overview of the tac 2010 knowledge base population track. In *Proceedings of the Third Text Analysis Conference*.

S.S. Kataria, K.S. Kumar, R. Rastogi, P. Sen, and S.H. Sengamedu. 2011. Entity disambiguation with hierarchical topic models. In *Proceedings of KDD*.

Jie Liu, Ruifeng Xu, Qin Lu, and Jian Xu. 2012. Explore chinese encyclopedic knowledge to disambiguate person names. In *The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2012)*.

R. Mihalcea and A. Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM.

Zehuan Peng, Le Sun, and Xianpei Han. 2012. Sinerd: A chinese named entity recognition and disambiguation system using a two-stage method. In *The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2012)*.

FAN Qing-hu, ZAN Hong-ying, CHAI Yu-mei, JIA Yu-xiang, and NIU Gui-ling. 2012. Chinese personal name disambiguation based on vector space model. In *The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2012)*.

Wei Tian, Xiao Pan, Zhengtao Yu, Yantuan Xian, Xiuzhen Yang, Yu Qin, and Wenxu Long. 2012. Chinese name disambiguation based on adaptive clustering with the attribute features. In *The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2012)*.

Longyue Wang, Shuo Li, Derek F. Wong, and Lidia S. Chao. 2012. A joint chinese named entity recognition and disambiguation system. In *The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2012)*.

Han Wei, Liu Guang, Mao Yuzhao, and Huang Zhenni. 2012. Attribute based chinese named entity recognition and disambiguation. In *The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2012)*.

Tao Zhang, Kang Liu, and Jun Zhao. 2012. The nlpr entity linking system at clp 2012.

Hao Zong, Derek F. Wong, and Lidia S. Chao. 2012. A template based hybrid model for chinese personal name disambiguation. In *The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2012)*.