Explore Chinese Encyclopedic Knowledge to Disambiguate Person Names

Jie Liu* Ruifeng Xu* Qin Lu† Jian Xu†
Key Laboratory of Network Oriented Intelligent Computation,
Shenzhen Graduate School, Harbin Institute of Technology, China*
{lyjxcz, xuruifeng.hits}@gmail.com

Department of Computing, Hong Kong Polytechnic University, Hong Kong†
{csluqin, csjxu}@comp.polyu.edu.hk

Abstract

This paper presents the HITSZ-PolyU system in the CIPS-SIGHAN bakeoff 2012 Task 3, Chinese Personal Name Disambiguation. This system leveraged the Chinese encyclopedia Baidu Baike (Baike) as the external knowledge to disambiguate the person names. Three kinds of features are extracted from Baike. They are the entities' texts in Baike, the entities' work-of-art words and titles in the Baike. With these features, a Decision Tree (DT) based classifier is trained to link test names to nodes in the NameKB. Besides, the contextual information surrounding test names is used to verify whether test names are person name or not. Finally, a simple clustering approach is used to group NIL test names that have no links to the NameKB. Our proposed system attains 64.04% precision, 70.1% recall and 66.95% F-score.

1 Introduction

With the development of the Internet and social network, more and more personal names appear on the web. However, many people share the same namesake, thus causing name ambiguities in online texts. A useful approach for disambiguating the person names is of great benefit to the information extraction and other natural language processing problems.

Worse still, Chinese personal name disambiguation is much more challenging. This is because it is difficult to locate the boundaries for Chinese personal names. In example 1,

Both "朱方勇/ZhuFangyong" and "朱方/ZhuFang" can be identified as named entities since the word "勇/Yong" (meaning "bravely") can be placed together with the word "闯/pass" to form a phrase.

Example 1: **朱方勇**闯三关 (ZhuFangyong passed three barriers)

In addition, some Chinese surnames are a combination of parents' family names. Take "张包子俊/Zhang-Bao Zijun" for example, the surname "张包/Zhang-Bao" was made by combining two signal-syllable family names "张/Zhang" and "包/Bao". This combination also makes the situation more complex. Moreover, some person names are simply common words. For example, "白雪/BaiXue" can refer to "white snow" when it doesn't refer to a person.

In recent years, many researches have been conducted on person name disambiguation. Web People Search (Artiles et al., 2009, 2010) provides a benchmark evaluation competition. In this task, a lot of approaches resolve personal name ambiguity by clustering approaches. Disambiguating personal names generally involved two steps: feature extraction step and document clustering. In terms of extracted features, Bagga et al. (1998) used the within-document coreference approach to extract the most relevant context for test names. Xu et al. (2012) added the key phrases as the features. Other researchers have also used URLs, title words, ngrams, snippets and so on (Chen et al., 2009; Ikeda et al., 2009; Long and Shi, 2010). To group text documents into different clusters, Hierarchical Agglomerative Clustering (HAC) is commonly used. Gong et al. (2009) proposed a method to train a classifier to select the best HAC cutting point. Yoshida et al. (2010) used a two-stage clustering by bootstrapping to improve the low recall val-

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 138–145,
Tianjin, China, 20-21 DEC. 2012

ues created in the first stage. Besides, some researchers incorporated the social networks of the test names to do person name disambiguation. Tang et al. (2011) established a bipartite graph by extracting named entities that co-occur with the test names and then resolute the person name ambiguity based on graph similarity. Lang et al. (2009) proposed to extend the social networks by using the search engine to achieve a better performance.

Similarly, the TAC-KBP entity-linking task has been held four times (McNamee et al. 2009, Chen et al. 2010, Zhang et al. 2011, Xu et al. 2012). A general architecture consists of three modules: candidate generation, candidate selection, NIL entities clustering.

In the candidate selection step, some researchers viewed it as an information retrieval task. Varma et al (2010) ranked the candidates with a TF-IDF weighting scheme. Fern et al. (2010) used the PageRank approach to rank the entities. Zhang et al. (2010) proposed a compound system by using the Lucene-based ranking, SVM-rank and binary SVM classifier. To rank the candidates, different features are used. Zhang et al. (2011) used surface features, contextual features and semantic features. In addition, they calculated the contexts' probability distribution over the Wikipedia categories to measure the topics' similarity. Chang et al. (2010) extracted anchor text strings as features. Lehmann et al. (2010) and Mcnamee (2010) utilized the Wikipedia links. In our system, a Decision Tree classifier has been used with four kinds of features: the entities' texts in NameKB, the entities' texts in Baike, the entities' work-of-art words and titles in the Baike.

Some test name may have no corresponding links to the entities in the knowledge base (KB) and will be classified as NIL queries. To detect these NIL queries, Chen et al. (2010) simply marked the queries without candidate as NIL. Lehmann et al. (2010) trained a classifier to find NIL queries.

Similar to the WePS and TAC-KBP tasks, the CIPS-SIGHAN CLP2012 bakeoff task was held to promote the Chinese personal name disambiguation. In this task, our system leveraged the Chinese encyclopedia Baidu Baike (Baike) as the external knowledge to disambiguate the person names, resulting in an F1 score of 66.95%.

The rest of the paper is organized as follows. Section 2 describes person name disambiguation task. Section 3 presents the design and implementation of our system for this task. Section 4 gives the performance achieved by our system. Section 5 gives the conclusion and future work.

2 Task Description

CIPS-SIGHAN bakeoff on person name disambiguation is an Entity-Linking task. In the task, 32 test names and a document collection for each test name are provided. Each document contains at least one test name mention. NameKB is also provided to describe entities related to the test name. Each entity with the short description is about one person in reality.

The systems are required to link documents to the corresponding entities in NameKB. Some test names are not named entities but common words. Documents containing these test names should be classified as "other". Other test names that cannot be linked to the NameKB are required to be clustered.

3 Person Name Disambiguation System

In our system, disambiguating personal names is conducted in five steps. In the first step, some preprocessing work will be done, for example, getting the information from encyclopedia, establishing one-to-one mapping between entities in Baike and in NameKB. In the third step, we will link test names mentions in documents to the entities in NameKB. As there is just a short description for each entity in the NameKB, we proposed to enrich the entities' description text by using four kinds of information. Finally, a DT based classifier trained was used to determine which result should be adopted. Then, documents in which the test name mentions have no linking to the entities in NameKB were decided whether their test name mentions refer to some person or just are the common words. In this common words identification step, the test names were judged whether there are the words describing people around them. Finally, simple clustering for the NIL documents was done by considering whether the words set around the test name mentions were sharing the words describing people.

3.1 Preprocessing

In order to use the rich information of the encyclopedia in the Baike, the 32 pages referring to the 32 test names are downloaded for the Internet. In each page, there are several subpages referring to same number of entities. As the Table 1 shown blow, there are 16 entities for the test name "白雪/BaiXue". For each subpage, there is rich information about the corresponding entity. We extract entities' titles, entities' contents and the entities' work-of-art names. In addition, all the texts used in our system are segmented.

1.歌手/singer
2.演员/actor
3.运动员/athlete
4.配音演员/dubbing speaker
5.画家/painter
6.作家/writer
7.《海豚湾恋人》插曲/interlude song of *Love at Dolphin Bay*
8.snowhite 文具/ stationery
9.小说《大秦帝国》女主角/heroine of the novel named *The Qin empire*
10.动漫人物/ cartoon character
11.布袋戏人物/ glove puppetry character
12.《活佛济公》角色/role in *The Legends of Ji Gong*
13.柯南主题曲/the theme song of Conan
14.南京籍演员/actor born in Namjing
15.《金陵十三钗》演员之一/one of the actor in *The Flowers of War*
16.汉语词汇/ word in Chinese

Table 1: Titles of entities in page describing person "白雪/BauXue"

3.2 Map the Baike to the NameKB

Though the various kinds of information were extracted from the Baike, we cannot directly use them in the task because we don't know which entity the information belongs to. In order to solve this problem, the one-to-one mapping between entities in Baike and entities in NameKB is established. For most test names the number of entities in Baike is bigger than the one in NameKB. But it is not always true for all test names that the entity set in Baike contains all the entities in the NameKB.

In this step, VSM is used to represent the entities' contents in both Baike and NameKB. The the nouns found in all the contents are selected as the features and weighted with the TF-IDF score. We then use the cosine metric as similarity calculation function.

It is not simply to select the most similar entity in NameKB for a given entity in Baike. We also must select the most similar entity in Baike for a given entity in NameKB to make the mapping be one-to-one. After establishing the mapping the additional entities both in Baike and in NameKB will be discarded. In the training dataset, this simple method gets the very higher precision.

3.3 Entity-Linking System

In this section, the entity-linking method is described. Entity-Linking system links the documents to the entities in NameKB. Our entity-linking method is a compound one. We built four entity-linking sub-systems by using different kinds of information. Each system gives an entity-linking result. The machine learning method is trained to get a classifier which will help us do better decision with the four entity-linking results given by the sub-systems.

The four entity-linking subsystems (S1, S2, S3 and S4) are described separately.

S1. Using the entity content in NameKB

In the NameKB, a short description is given for each entity. In this subsystem, the similarity between the descriptions in NameKB and the documents in collections was measured to determine whether there is a link between them. In this subsystem, a vectorial representation of document with the test name is compared with the vectorial representations of the entities' descriptions in NameKB. The features used in these vectorial representations are all nouns with assigned TF-IDF scores. The subsystem chooses the NameKB entity which has the maximum similarity with the document as the output. The threshold for the minimum similarity value is set empirically to get the higher accuracy. The documents with similarity being less than a given threshold (0.27 in this task) will be classified as NIL queries, indicating that they have no link to the entities in NameKB.

S2. Using the entity content in Baike

There is richer information in the Baike than in the NameKB. Baike has information box, events list, work-of-art words and so on. These are very useful to disambiguate the test names. Like the S1 subsystem, the similarity between the entities' contents in Baike and the documents in collections was measured to get the most similar entity for each document. The threshold for the minimum similarity value is set empirically, too. Like the S1, the documents less than the given threshold (0.15) will be classified as NIL queries. The result is intermediate one. Then, it is used as the input to get the final result by leveraging the mapping established in 3.2.

S3. Using the work-of-art name string in Baike

The entities in the NameKB are mostly famous person, such as artists, government officials, authors, actors, singers, researchers and so on. There are a lot of work-of-art names marked as "《" and "》" in their descriptions. These names are the names of books, songs, movies, conferences, journals and so on. In most cases we can identify which entity the test name mentioned in a document refers to. It is difficult to make decision when there are more than one entities sharing the same work-of-art names, for example, "EI" is shared by many professors. In order to avoid misjudging in that case, duplicates are removed to get the work-of-art names lists for each entity.

Because most of the work-of-art names will be segmented into several words, we avoid this issue by directly looking up the name strings in each document. The farther away from the test names, the less relevant to them. Based on that observation the boundary for looking up is set to get the better result. Our system just looks up the string names in the substrings containing the test names. The looking up windows is set as 40 characters centered in the test names. If finding, the document will be marked with the corresponding entity. This result is also the intermediate one. Mapping will be done to get the final one. The documents in which the name strings ware not found will be marked with a special tag.

S4. Using the entity title in the Baike

In the Baike, for each entity there is a title to give a very short and exact description, such as "柔道运动员/judo artist", "南京大学副教授/associate professor of Nanjing University". With these short titles we can get some very useful information about the entities. For example we can get entities' organizations, occupations and so on. In this subsystem, the ending words of the titles are used only since for most titles the ending words are the occupations of the entities. We just simply look up the occupation words extracted from the titles in the documents. Similar to the S3 subsystem, the looking up boundary is set to get the better result. The mapping the intermediate result to the final one is also needed.

From four subsystems described above, we get four results which tell us how to link the documents in the collections to the entities in NameKB. In order to combine these results, machine learning method is used to get the best final result. With the training set, a DT based classifier is trained. Features for the DT classifier is shown blow in Table 2. For example, the value S1 will be Y if the subsystem S1 finds a link between the document and some entity in NameKB. Otherwise, N will be assigned to it if S1does not find a link for the document. The value for S12 is if the subsystems S1 and S2 both find the same link for the document. Similarly, the value of feature S1234 indicates whether the four subsystems S1, S2, S3, S4 find the same link for the document. Five classes are trained for classification. They are shown in Table 2. This classifier is applied to determine which result should be adopted.

Feature	Value
S1,S2,S3,S4	Y: find a link by Si N: find not link by Si
S12,S13,S14,S23,S24,S34	Y: find the same link by Si and Sj N: other
S123,S124,S134,S234	Y: find the same link by Si, Sj and Sk N: other
S1234	Y: find the same link by S1, S2, S3 and S4 N: other

Table 2: The features in the DT classifier

Classes	Remark
AS1, AS2, AS3, AS4	Find the link and the result of Si is adopted
N	There is not link

Table 3: Five classes in the DT classifier

For each document in test set, the four subsystems give four results. The classifier trained in training set tells which subsystems' result should be adopted. For example, some document is labeled by the classifier as the S2, which means the classifier tells us that the link is found and the result of S2 (by using the entities' text in Baike) should be adopted. The documents which are classified in the class N are told that there is no corresponding entity in NameKB.

3.4 Identifying Common Words

The test name words (the words exactly matching the test names and mentioned in the documents) do not always refer to person or named entity. In some documents they are common words. For the test name “白雪/BaiXue”, in Example 1, “白雪/BaiXue” is a person name and refers to a marathoner while in Example 2, “白雪/BaiXue” is a common words meaning “white snow” rather than a person name.

Example 1: 白雪获女子马拉松冠军(BaiXue won the women's marathon champion)
Example 2: 海拔 5100 米的玉树雪山披着白雪 (The Yushu snow mountain with the altitude of 5100 meters is covered with white snow)

In this task, the systems are required to find out these common words and to mark them as “other”. But in the NameKB of the training set, some test names have the common word entities, such as “黄海/HuangHai”, ”黄河/HuangHe”, “华山/HuaShan”, “华明/HuaMing”, “方正/FangZheng” and so on. And the documents referring to these common word entities were marked as the entities numbers rather than “other”. So “other” is only be labeled on the documents in which the test names don’t refer to the entities in NameKB and refer to common words. Base on that observation, our system just identify whether the test names are the common words after entity linking. That means the common words identification is just for those documents which have no links to the NameKB entities.

In this step, the words surrounding the test names within a given window size are collected to identify the common words. If the surrounding words contain person names or occupations, the test names will be identified as the person name. Otherwise, test names will identified as common words and the corresponding document will be marked with “other”.

Take the test name “丛林/LinCong” for example, in example 3, the surrounding word set is {“流沙/ShaLiu”, “李世荣/ShirongLi”, “毋巨龙/JulongWu”, “王珍祥/ZhenxiangWang”} when the window size is set to 2 noun. In the word set, “李世荣/ShirongLi”, “毋巨龙/JulongWu”, and “王珍祥/ZhenXiangWu” are person names, but “流沙/ShaLiu” is not recognized as person name by the POS tagging tools. So the document document is expected to refer to some people. In the Example 4, because the surrounding word set {“厅/department”, “厅/director”, “印花/print”, “基地/base”} contains {“厅长/director”}--an occupation word, the test name string in the document will also denote a person. In the Example 5, the corresponding document will be marked as “other” because the test name mention’s surrounding word set {“两岸”, “峰峦”, “河道”, “水流”} contains neither person name nor occupation word. A simple dictionary-based occupation word identification is developed in this step
Example 3: 【作者】陈亮；流沙；李世荣；丛林；毋巨龙；王珍祥； (Authors: Liang Chen, Shan Liu, Shirong Li, Lin Cong, Julong Wu, Zhenxiang Wang)

Example 4: 福建省科技厅厅长丛林来访“冷转移印花示范基地” (Lin Cong, the director of the Science and Technology Department of Fujian Province, visited the cold transfer printing model base)
Example 5: 两岸峰峦竞秀，丛林密布，河道曲折迂回，水流缓急有致 (River twists and turns across the rising mountains which are covered with dense jungles)

After this step, the documents in which the test name mentions are just the common words will be selected and marked as “other”. All other documents will be clustered in next section.

3.5 NIL Document Clustering

The documents without the mark “other” are required to be clustered together based on the underlying entities.

In our system, a simple clustering is done among these documents. The words around the test names within a certain window (2 words) are collected as the documents’ words sets. All the person words (person names and the occupations words) in the words sets are picked upchosen to measure whether these documents share the same person wordshave words in common. If so, The the documents share the same person words will be clustered togethergrouped into clusters.

For the test name “李晓明/XiaomingLi”, because the doc405 and the doc332 will be grouped since they have the same word share the person name “董事长/chairman”, they will be clustered together. For the test name “李晓明/XiaomingLi”, because the doc405 and the doc332 share the person name “董事长/chairman”, they will be clustered together.
Doc 405 : 秦/Qin 龙/Long （ 国际/international ） 集团/Group 董事长/chairman 李晓明/LiXiaoming 到/go to 黑龙江/Heilongjiang Province 交通职业技术学院

/Communication Polytechnic college 参观/visit 考察/inspect
Doc 332: 市委/ municipal Party committee 书记 /secretary 杨信/XinYang 陪同/together 北京 /Beijing 秦/Qin 龙/Long 国际/international 公司 /company 董事长 /chairman 李晓明 /XiaomingLi 一/one 行/coming 来到/go to 扎龙 /Zhalong
Doc 405: personal words set { "董事长 /chairman"}
Doc 332: personal words set { "董事长 /chairman"}

4 Performance Evaluations

This section shows evaluations of our system for the CIPS-SIGHAN bakeoff 2012 Task 3 in training set and the final test set. The results are shown in Table 4.

Data set	Precision	Recall	F1
Training set	0.6761	0.7277	0.7010
Test set	0.6404	0.7013	0.6695

Table 4: The performance of our system

It is shown that our system achieves the higher recall performance than the precision. In addition, the result on the training set is higher than the one on the testing set both in the precision and recall.

To validate the usefulness of the leveraging the encyclopedia, we conducted an experiments with and without using the encyclopedia. Experimental result in Table 5 shows that leveraging the encyclopedia Baike gives remarkable improvement.

Runs	Precision	Recall	F1
Without Baike	0.6399	0.5973	0.6179
With Baike	0.6761	0.7277	0.7010

Table 5: Performance evaluation by leveraging the Encyclopedia Baike

In addition, three sets of experiments are conducted separately on the training dataset to measure the effectiveness of our system in entity linking, common word identification, and document clustering. They are denoted as PureEL, PureCWI and PureCluster, respectively. In the golden answer of the training dataset, there are three types of documents: documents that can be linked to the NameKB, documents that are classified as "other" and documents which are categorized as "NIL" for clustering. PureEL simply considers documents that can be linked to nodes in the NameKB. Our system evaluates the performance in linking these documents to the NameKB in Table 6. Experimental results show that our system achieves a high precision (87.5%) and F-score (82.3%) in linking documents to nodes in NameKB.

PureCWI takes into account documents that are classified as "other" and "NIL" categories in the golden answer for training dataset. Documents of "NIL" categories are introduced as noises to testify the robustness of our system in identifying names as common words. Experimental results in Table 6 indicate a high recall but at the cost of low precision, implying that documents of "NIL" categories affect the performance of common word identification.

PureCluster simply uses the documents of "NIL" categories. Results in Table 6 shows our system achieves a high precision in clustering documents, indicating that our system introduces less noise in clustering solutions. However, our system has a low recall in clustering, implying that the number of clusters produced by our system is less than that of the manually assigned categories in the golden answer. Through further analysis, we found that most of documents of "NIL" categories are placed into a singleton clusters.

Runs	Precision	Recall	F1
PureEL	0.875	0.777	0.823
PureCWI	0.231	0.762	0.355
PureCluster	0.917	0.456	0.609

Table 6: The performance data of the subsystems

5 Conclusions and Future Work

The HITSZ-PolyU system enriches the information of the entities in given NameKB by leveraging the encyclopedia Baike. Experiments have shown that it is very helpful in the task. For the entity linking, four results are got by using different information. A DT based classifier was used to combine the four results to get the final one. A simple approach to predict whether the test name mentions is common words is used but not very useful. More powerful common words identification method will be considered to get better performance. The words matching based clustering does achieve the good performance.

Better clustering approach should be applied to improve the performance. In addition, the using of the Baike in our system is very simple. The new way how to make better use of it should be considered in the future researches. Furthermore, in mapping establishing step the additional entities in Baike was discarded directly. However, those additional entities should be used before the clustering step to filter out the documents which has the link to them, which can alleviate the clustering problem.

References

Amit Bagga, Breck Baldwin. 1998. Entity-Based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* v.1, pp: 79-85.

Angel X. Chang, Valentin I. Spitkovsky, Eric Yeh, Eneko Agirre and Christopher D. Manning. 2010. Stanford-UBC Entity Linking at TAC-KBP. In *Poceedings of the Third Text Analysis Conference (TAC 2010).*

Chong Long and Lei Shi. 2010. Web person name disambiguation by relevance weighting of extended feature sets. In *Third Web People Search Evaluation Forum (WePS-3), CLEF 2010.*

Ergin Elmacioglu, Yee Fan Tan, Su Yan, Min-Yen Kan and Dongwon Lee. 2007. PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features. In *Proceedings of the Second Text Analysis Conference (TAC 2007).*

Elena Smirnova, Konstantin Avrachenkov, and Brigitte Trousse. 2010. Using web graph structure for person name disambiguation. In *Third Web People Search Evaluation Forum (WePS-3), CLEF 2010.*

Javier Artiles, Julio Gonzalo, 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, 2009.*

Javier Artiles, Andrew Borthwick, Julio Gonzalo, Satoshi Sekine, and Enrique Amigo. 2010. WePS-3 evaluation campaign: overview of the web people search clustering and attribute extraction tasks. In *Third Web People Search Evaluation Forum (WePS-3), CLEF 2010.*

Jian Xu, Qin Lu, Jie Liu, Ruifeng Xu. 2012. NLP-Comp in TAC 2012 Entity Linking and Slot-Filling. In *Proceedings of the Fourth Text Analysis Conference (TAC 2012).*

Jian Xu, Qin Lu, Zhengzhong Liu. 2012. Combining classification with clustering for web person disambiguation. *In Proceedings of the 21st International Conference Companion on World Wide Web*, pp: 637-638.

Jintao Tang, Qin Lu, Ting Wang, Ji Wang, and Wenjie Li. 2011. A Bipartite Graph Based Social Network Splicing Method for Person Name Disambiguation. In *SIGIR 2011*, pp. 1233-1234.

John Lehmann, Sean Monahan, Luke Nezda, Arnold Jung and Ying Shi. 2010. LCC Approaches to Knowledge Base Population at TAC 2010. In *Proceedings of the Third Text Analysis Conference (TAC 2010)*.

Jun Gong, Douglas W. Oard. 2009. Selecting hierarchical clustering cut points for web person-name disambiguation. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp: 778-779.

Jun Lang, Bing Qin, Wei Song, Long Liu, Ting Liu, Sheng Li. 2009. Person Name Disambiguation of Searching Results Using Social Network. *Chinese Journal of Computers*, No.7, pp: 1365-1374.

Krisztian Balog, Jiyin He, Katja Hofmann, Valentin Jijkoun, Christof Monz, Manos Tsagkias, Wouter Weerkamp and Maarten de Rijke. 2009. The University of Amsterdam at WePS2. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.*

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, 2009.*

Minoru Yoshida, Masaki Ikeda, Shingo Ono, Issei Sato, Hiroshi Nakagawa. 2010. Person name disambiguation by bootstrapping. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp: 10-17.

Norberto Fernández, Jesus A. Fisteus, Luis Sánchez, and Luis Sánchez. 2010. WebTLab: A cooccurrence-based approach to KBP 2010 Entity-Linking task. In *Proceedings of the Third Text Analysis Conference (TAC 2010)*.

Paul McNamee, Mark Dredze, Adam Gerber, Nikesh Garera, Tim Finin, James Mayfield, Christine Piatko, Delip Rao, David Yarowsky and Markus Dreyer. 2009. HLTCOE approaches to knowledge base population at TAC 2009. In *Proceedings of the Second Text Analysis Conference (TAC 2009).*

Paul McNamee. 2010. HLTCOE Efforts in Entity Linking at TAC KBP 2010. In *Proceedings of the Third Text Analysis Conference (TAC 2010)*.

Sanyuan Gao, Yichao Cai, Si Li, Zongyu Zhang, Jingyi Guan, Yan Li, Hao Zhang, Weiran Xu and Jun Guo. 2010. PRIS at TAC2010 KBP Track. In *Proceedings of the Third Text Analysis Conference (TAC 2010).*

Vasudeva Varma, Praveen Bysani, Kranthi Reddy, Vijay Bharath Reddy, Sudheer Kovelamudi, Srikanth Reddy Vaddepally, Radheshyam Nanduri, Kiran Kumar N, Santhosh Gsk and Prasad Pingali. 2010. IIIT Hyderabad in Guided Summarization and Knowledge Base Guided Summarization Track. *In Proceedings of the Third Text Analysis Conference (TAC 2010)*.

Wei Zhang, Jian Su, Bin Chen, Wenting Wang, Zhiqiang Toh, Yanchuan Sim, Yunbo Cao, Chin Yew Lin and Chew Lim Tan. 2011. I2R-NUS-MSRA at TAC 2011: Entity Linking. *In Proceedings of the Fourth Text Analysis Conference (TAC 2011)*.

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, 2009.*

Zheng Chen, Suzanne Tamang, Adam Lee, Xiang Li, Wen-Pin Lin, Matthew Snover, Javier Artiles, Marissa Passantino and Heng Ji. 2010. CUNY-BLENDER TAC-KBP2010 Entity Linking and Slot Filling System Description. In *Proceedings of the Third Text Analysis Conference (TAC 2010).*