Chinese Personal Name Disambiguation Based on Vector Space Model

Qing-hu Fan
College of Information Engineering,
Zhengzhou University, Zhengzhou,
Henan ,China
fanqinghude@163.com

Hong-ying Zan Yu-mei Chai
Yu-xiang Jia
College of Information Engineering,
Zhengzhou University, Zhengzhou,
Henan ,China
{iehyzan, ieymchai, ieyx-jia}@zzu.edu.cn

Gui-ling Niu
Foreign Languages School,
Zhengzhou University, Zhengzhou,
Henan ,China
mayerniu@163.com

Abstract

This paper introduces the task of Chinese personal name disambiguation of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP) 2012 that Natural Language Processing Laboratory of Zhengzhou University took part in. In this task, we mainly use the Vector Space Model to disambiguate Chinese personal name. We extract different named entity features from diverse names information, and give different weights to various named entity features with the importance. First of all, we classify all the name documents, and then we cluster the documents that cannot be mapped to names that have been defined. Eventually the results of classification and the clustering are combined. In the test corpus experiments, the accuracy rate is 0.6778, the recall rate is 0.7205 and the F value is 0.6985 for all names.

1 Introduction

Named Entity is the fundamental information elements in text, and is the basis for understanding the text correctly. Named Entities include person names, organization names, place names, time, date, and digital. Named Entity Recognition is to identify the entities in the text and determine what category it is. Such as 方正 fangzheng ‘Fang Zheng’, maybe the name is an associate professor at the Department of Mechanical and Electrical Engineering, Physics and Electrical and Mechanical Engineering College of Xiamen University, or it may be Peking University Founder Group Corp that was established by Peking University. It needs to associate with context for disambiguating the entity Fang Zheng. For example, Fang Zheng who is an associate professor at Xiamen University can be extracted with the feature that Xiamen University, Mechanical and Electrical Engineering College or associate professor, which can eliminate ambiguity.

2 Related Research

In the early stages of Named Entity Disambiguation, Bagga and Baldwin (1998) use Vector Space Model to resolve ambiguities between people having the same personal name. Han and Zhao (2010) proposed a knowledge-based method that captures structural semantic knowledge in multiple knowledge sources to disambiguate personal entities. Han and Sun (2011) proposed a generative Entity-Mention model that leverages heterogeneous entity knowledge for the entity linking task. In Chinese person name disambiguation, Li, et al (2010) carried out the first conference, Chinese Language Processing (CLP-2010), which contains Chinese person names disambiguation task. In this task Shi, et al (2011) proposed a post-processing method that is based on multiple entity recognition system integration

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 152–158,
Tianjin, China, 20-21 DEC. 2012

and heuristic rules, Zhang, et al (2010) proposed a method that extracts various person features to identify different person names, and according to the Chinese word segmentation, we constructed artificially rules that identify the names correctly. We propose a method that is based on various entities recognition and initialize evaluation for the features that are the common characteristics of different names, and then take Vector Space Model to calculate it. In the end, the documents that cannot be mapped to names that have been defined in the knowledge base are clustered into different types.

CLP2012 Named Entity disambiguation is a task pre-classification and later clustering problems. The task provides a knowledge base of Chinese names which include multiple definitions of personal names, and some documents about person names. It is the purpose of the task that makes each name that appears in documents to link corresponding definition of the knowledge base, and makes the documents that cannot link to corresponding definition of the knowledge base to cluster which two documents have the same named Entity feature. Task input: Names knowledge base of named Entity, text set corresponding each name. Task output: if the name of each text links to the knowledge base of a definition, then output the corresponding id, if the name of each text is ordinary words, then output "other", if the name of each text does not belong to the above two kinds, then output Numbers: Out_XX that have been put into.

This paper is organized as follows: in section 3 we will introduce the method that extract the named entities related to figures. In section 4 we will introduce the calculation model of the named entities. In section 5 we will describe experiments and results. In the last section we will make conclusions and future work.

3 Extract the Named Entities Related to Figures

3.1 Character works

Works have the originality and are the intellectual creations that can be copied in a certain physical form in the field of literature and science.[1] Works include literature works, music, drama, folk art forms, dance works, photographs, films, television, video works, etc. Character works is the significant characteristic to identify figures. In evaluation corpus, it is generally that a character works correspond to one specific character. Therefore, it is character works that plays an important role to eliminate name disambiguation.

Extraction method: we extract character works from each figure corpus; in other words, we extract all the contents of quotation marks.

Format the character works:

1) If there is 之 Zhi that appears in the work, then we split the work with 之 Zhi.

For example:

白云(孙皓晖先生的长篇小说《大秦帝国之黑色裂变》中所虚构的女主角)

Bai-Yun(Sun-hao-hui-xian-sheng-de-chang-pian-xiao-shuo-da-qin-di-guo-zhi-hei-se-lie-bian-zhong-suo-xu-gou-de-nv-zhu-jiao)

Bai-Yun(she is the fictional actress in the Danqin Empire with The Black Fission that is Mr.Sun Haohui's novel)

We will extract 大秦帝国之黑色裂变 da-qin-di-guo-zhi-hei-se-lie-bian 'Danqin Empire with The Black Fission' that is the work, however the work cannot be identified. As 大秦帝国之黑色裂变 da-qin-di-guo-zhi-hei-se-lie-bian 'Danqin Empire with The Black Fission' is only the first novel of 大秦帝国 da-qin-di-guo 'Danqin Empire' in literature works.[2] We split 大秦帝国之黑色裂变 da-qin-di-guo-zhi-hei-se-lie-bian 'Danqin Empire with The Black Fission' into 大秦帝国 da-qin-di-guo 'Danqin Empire' and 黑色裂变 hei-se-lie-bian 'The Black Fission'with 之 Zhi, and then they can be identified correctly.

2) If the length of works' name is less than 2, it is required to extract works and quotation marks.

Eg:马啸担任河南卫视《旅游》栏目主持人. Ma-xiao-dan-ren-hen-nan-wei-shi-lv-you-lan-mu-zhu-chi-ren 'Ma Xiao is appointed host of Traveling program in Henan TV'In this sentence 旅游 lv-you 'Traveling' is the work name. It is known that Traveling has different part of speech, which can be a verb or noun. The Traveling is a TV program in the sentence, which is a noun. It will reduce accuracy rate that we take Traveling as the feature.

3.2 Character Aliases

Aliases are the names other than the formal or specific. They are used in writing, oral.[3] Character aliases are an essential feature for eliminating

[1] http://baike.baidu.com/view/94574.htm

[2] http://baike.baidu.com/view/525001.htm

[3] http://baike.baidu.com/view/343250.htm

the disambiguation. We define that each file-name in KB folder is the figure's original name, others are character aliases. We use the methods that are based on pattern matching to extract character aliases Lu and HOU (2006), as is shown below following methods:

1) Synonymy keywords + Synonyms + End identifier

Synonymy keywords: 本名｜别号｜，字｜^(字)｜，号｜^(号) ｜又号｜^(名)｜笔名｜自号｜又名｜乳名｜别名｜原名｜艺名｜本名｜曾用名｜俗称｜亦称｜又称, the symbol "|"means choose, "^"means that matches the beginning of the string.

End identifier: it means the end of extracting the synonyms, the end signs are always (，or,) and (。or .), which mean comma symbol and full stop. If we extract character aliases equal with original names, then we should use the feature that synonyms combine with synonymy.

Eg: 白云（原名杨维汉，广东省潮安县人）Bai-Yun(Yuan-ming-yang-wei-han-guang-zhou-chao-an-xian-ren) Bai-Yun(Her original family name is Yang Weihan and she was born in ChaoAn Guangdong Province).

According to the first method that we could extract 杨维汉 yang-wei-han 'Yang Weihan' that it is character aliase. However, the content of 白雪 bai-xue 'Bai Xue' that 白百何，中国内地女演员，别名白雪 Bai-bai-he-zhong-guo-nei-di-nv-yan-yuan-bie-ming-bai-xue 'Bai baihe is Chinese mainland actress and her aliase is Bai Xue' and 陈大威，号白雪，碧松斋主人 chen-da-wei-hao-bai-xue-bi-song-zhai-zhu-ren 'Chen Dawei's art-name is Bai Xue and he is the host of Bi-Song-Zhai', we could extract 白雪 Bai Xue that it is character aliase, which we cannot make a distinction between the two characters. As a result we take 别名白雪 bie-ming-bai-xue 'aliase is Bai Xue' and 号白雪 hao-bai-xue 'art-name is Bai Xue' as the features to eliminate disambiguation.

2) (Original family name|^(Chinese surnames))+ name+ end identifier

Original family name: we take original family name as prefix.

^ (Chinese surnames): it means the beginning of the Chinese; Zhang, et al (2008) found out that the top 400 Chinese surnames have covered 99%.

End identifier: it is the same define as the first method.

If the length of character aliases are less than 2 or more than 3, and then they will be extracted.

Eg1: the content of 白雪 Bai Xue that 白百何，中国内地女演员，别名白雪 Bai-bai-he-zhong-guo-nei-di-nv-yan-yuan-bie-ming-bai-xue 'Bai baihe is Chinese mainland actress and her aliase is Bai Xue' in the sentence the family name of Bai Xue is Bai. End identifier is "," ,then we could extract "白百何" as character aliase from the first method.

Eg2: the content of Baixue that 陈大威，号白雪，碧松斋主人 chen-da-wei-hao-bai-xue-bi-song-zhai-zhu-ren 'Chen Dawei's art-name Bai Xue and he is the host of Bi-Song-Zhai', in this sentence the family name of Bai Xue is Bai,and we know that 陈大威(Chen Dawei) is character aliase, the family Bai is different from 陈 Chen. Therefore, according to second method we use the family name Chen. End identifier is "," ,then we extract character aliase as Chen Dawei.

3.3 Named Entity

Named Entity is the feature to discriminate figures. The features related to figure, Learning Unit, organizations, living space, and other entities, can mark different figures. In this task, we primarily extract features learning unit, organizations, living space, and other entities.

1) Learning unit

Learning unit include university and college.

Extraction rules: (prefix end identifier | ns) + University name+ (University| college)

Prefix end identifier: it means the prefix end identifier of extracting learning unit; the same methods are used in character aliases.

Ns: it means place name.

Extraction process is shown as the following:

First, we use Peking University participle software to segment the character information corpora Yu, et al (2002).

Second, in order to judge the beginning of string we add "#" to the beginning of each character definition.

Third, we index the keywords "University" or "college" in the corpora.

Fourth, it is the direction that university's or college's prefix to loop for each participle units.

Fifth, if the next participle units contain "ns" or "#", the loop will stop.

Sixth, get the Chinese string that is between the beginning and the end index.

2) Organization and other entities

We use "nt" to express organization, and use "nz" to express other entities Yu, et al (2002). Then the Chinese words contain "nt" and "nz" will be extracted.

3) Living space

We mainly extract the highest frequency Chinese words in participle information; the word frequency determines the related degree about figure.

3.4 Figure Title

Title is the name that is set up, which refers to marriage, social relations, the status, and occupation. Such as professor, chief, director, etc. Title can help to distinguish different profession and status, which is essential for distinguishing various figures.

The figure title resource is part of Hownet[4] in this task, which contains 240 titles. We delete 28 titles that they reduce accuracy rate from title resource and add 8 titles that increase accuracy rate as title resources. As is shown in table 1:

Type	Titles
Be Deleted	代表 演员 领导 教授 组长 记者 委员 主任 黄河 书记 主席 姑娘 居民 老人 朋友 亲属 学生 儿子 夫人 父亲 继母 母亲 小姑娘 毕业生 村民 分子 专家 学员
Be Added	歌手 副教授 副主任 配音演员 喜剧演员 影视演员 相声演员 快板演员

Table 1: The titles of be deleted and added

Finally, we get a title resource that contains 220 titles. We will extract the titles that appear in title resource and in figures' definition, which will be title features, or it will be null.

4 Calculation Model

4.1 Vector Space Model

Vector Space Model (VSM) is algebraic model for representing text documents as vectors of identifiers. It is using vectors of identifiers that greatly improve computability of documents. In VSM each document can be expressed as N-dimensional vectors of identifiers, each dimensional can be chosen keywords as vector, which is shown as the following:

$$D_T = < T_1, T_2, T_3, \cdots, T_n >$$

T_i represents the i[th] item in document. ω_i represents weight of T_i, which is shown as the following:

$$D_\omega = < \omega_1, \omega_2, \omega_3, \cdots, \omega_n >$$

4.2 Feature Weight Calculation

Feature weight is used to reflect the importance for feature item in the document. Originally we calculate feature weight with Boolean weight that if the feature appears in the document, then the feature weight is 1, otherwise 0. However, this calculation method cannot reflect the importance of feature, and then we use Term Frequency (TF) and Relative Word Frequency to calculate, TF is the method that get frequency of feature item. Relative Word Frequency refers to the TF-IDF method.

But owing to the fact that each character information text is short, the above three kinds of feature weight calculation methods cannot effectively reflect importance of different characters. According to the section 3 that there are seven character information features: character works, character aliases, learning unit, organization, other entities, living space, and character title, which face the different importance of character features, we initialize weight for each character features. λ_i represents weight of the i[th] character feature. Each document can be expressed as seven character features in the following:

$$D_\lambda = < \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7 >$$

Generally, the experience parameters are for: λ_1=10, λ_2=6, λ_3=3, λ_4=2, λ_5=2, λ_6=5, λ_7=3, we use two methods to disambiguate Chinese personal name.

1) Term Frequency (TF):

$$D_{Out\_num} = MAX\left\{ \overset{n}{\underset{num=1}{D}} \left\{ \sum_{i=1}^{7} \lambda_i \cdot TF_i \right\} \right\}, \quad (1)$$

In formula (1), D_{Out_num} presents the definition that the character id is num in each document. If $D_{Out\_num} = 0$, then the document presents other. $\sum_{i=1}^{7} \lambda_i \cdot TF_i$ represents product weight-sum that initial weight and absolute frequency. $\overset{n}{\underset{num=1}{D}}$ represents the num($1 \le num \le n$) for each defi-

[4] http://www.keenage.com/

nition in each character information, n represents the total number of id for each character.

2) Vectorial Angle Cosine

$$Sim(D_i, D_j) = \frac{\sum_{k=1}^{t} W_{ik} \cdot W_{jk}}{\sqrt{\sum_{k=1}^{t} (W_{ik})^2 \cdot \sum_{k=1}^{t} (W_{jk})^2}}, \quad (2)$$

In formula (2), t represents vector dimension of each document features. W_{ik} represents the k^{th} vector dimension weight of the document D_i.

4.3 Documents Clustering

We cluster the documents from the number results of section 4.2 are "other". The steps are shown as the following:

- We extract the documents from the classification number results of section 4.2 as "other".
- We extract the character features, character work, character aliases, learning unit, and character title, from the documents by using the same method in section three.
- Boolean weighting

If two documents have the same feature that it is one of all, we cluster the two documents to one kind; otherwise, the document corresponds to the classification number of "other".

- Merge the results of section 4.2 and the results of section 4.3.In other words, the results of section 4.3 replace the classification number of "other" of the results of section 4.2.

5 Experiments

5.1 Experimental Data

We use the texts in the training corpus and test corpus of CLP2012. There are 16 character names and 1634 documents in training corpus, and 32 character names and 5503 documents in test corpus. The corpus has two kinds:

1) Knowledge base of named entity

It will provide a knowledge base for each name. For example, the name Fang Zheng refers to 12 entities, some of them are shown below:

- Fang Zheng(Comedian)
- Fang Zheng(Peking University Founder Group Corp)
- Fang Zheng (Associate professor)

2) It will provide a text set for each Name

5.2 Evaluation Method

We still take Fang Zheng as an example. It is defined as 12 kinds of entity in a knowledge base. The test document set that contain Fang Zheng is T. The reference answer marks the texts that contain Fang Zheng:

There are kinds of definition for Fang Zheng in the knowledge base. Each definition belongs to a class, which is expressed as L_XX($01 \leqslant XX \leqslant 12$), "XX" represents the definition of the XX^{th} entity.

If Fang Zheng is not an entity name but a common word, it belongs to the class of "other".

Fang Zheng is an entity name, but it has no definition in the knowledge base, then it belongs to Out_XX, XX represents id. Out_XX represents respectively Out_01, Out_02…

We always assume that when Fang Zheng appears in a text many times and their mark is the same. Therefore, a text is only given a marked result. This system marks the results that contain Fang Zheng with SL_XX, SOther, and SOut_XX respectively, and each text is only marked by one class. Then we calculate the precision rate and recall rate for each text are as follows:

1) If Fang Zheng that includes t is divided to SL_XX, then it is taken as definition of the knowledge base to calculate precision rate and recall rate are as follows:

$$\Pr e(t) = \frac{|SL\_XX \cap L\_XX|}{|SL\_XX|}$$

$$\mathrm{Re}\, c(t) = \frac{|SL\_XX \cap L\_XX|}{|SL\_XX|}$$

2) If Fang Zheng that includes t is divided to SOther, it is taken as a common word to calculate precision rate and recall rate are as follows:

$$\Pr e(t) = \frac{|SOther \cap Other|}{|SOther|}$$

$$\mathrm{Re}\, c(t) = \frac{|SOther \cap Other|}{|SOther|}$$

3) If Fang Zheng that includes t is put into SOut_XX, but t belongs to Out_YY in reference answer, the precision rate and recall rate are as follows:

$$\Pr e(t) = \frac{|SOut\_XX(t) \cap Out\_YY(t)|}{|SOut\_XX|},$$

$$\mathrm{Re}c(t) = \frac{|SOut\_XX(t) \cap Out\_YY(t)|}{|SOut\_XX|}$$

1) For a name that it is Fang Zheng, and then the precision rate and recall rate are as follows:

$$\Pr e(Fang\ Zheng) = \frac{\sum_{t \in T} \Pr e(t)}{|T|}$$

$$\mathrm{Re}c(Fang\ Zheng) = \frac{\sum_{t \in T} \mathrm{Re}c(t)}{|T|}$$

2) For all names, the precision rate and recall rate are as follows:

$$\Pr e = \frac{\sum_{n} \Pr e(n)}{|N|}, \mathrm{Re}c = \frac{\sum_{t} \mathrm{Re}c(t)}{|N|}$$

$$F = \frac{2 \times \Pr e \times \mathrm{Re}c}{\Pr e + \mathrm{Re}c}$$

5.3 Experimental Results

We use two methods that Term Frequency (TF) and Vectorial Angle Cosine (VAC) to disambiguate Chinese personal name. Two methods results are shown in Table 2.

Method	Pre	Rec	F
TF	0.6399	0.6795	0.6590
VAC	0.5972	0.6079	0.6025

Table 2: The results of two methods

We can see that TF method is superior to Vectorial Angle Cosine (VAC) method from table 2. Therefore, we mainly use TF method to eliminate discrimination on test corpus. The results are shown as table 3:

Method	Pre	Rec	F
TF	0.6778	0.7205	0.6985

Table 3: The results of test corpus

First, we can see the recall rate of the test corpus is not ideal from table 3. The problem is that we cannot extract enough named entity features in the content. Such as company name, and verb structures, etc. Second, the precision rate is low. The problem is that the estimation of initial weight of each named entity features and the clustering algorithm.

6 Conclusions and Future work

In this task we extract different named entities features from diverse names information, and give different weights to various named entities features with the importance of various named entities. Firstly, we classify each name documents. Secondly, we cluster the documents that cannot be mapped to names that have been defined. Finally, the results of classification and the clustering are combined. However, it is only the experience weight for the estimation of initial weight of each named entity features, then different weights have different effects. The Boolean method cannot fully reflect the importance of all kinds of named entities features.

In the future, we can expand the named entity features, such as company name, verb structures, and the noun near character name in the documents. Then we choose more effective named entity initial weights, and use various clustering methods for character documents (Sun, et al2008).

References

Bagga, A. and Baldwin, B. 1998. Entity-Based Cross-Document Coreferencing Using The Vector Space Model. Proceedings of the 17th international conference on Computational linguistics-Volume 1, pp.79-85.

Han Xianpei and Zhao Jun. 2010. Structural Semantic Relatedness: A Knowledge-Based Method to Named Entity Disambiguation. In: Proceedings of the 49th ACL.

Han Xianpei and Sun Le. 2011. A Generative Entity-Mention Model for Linking Entities with Knowledge Base. In: Proceedings of ACL-HLT.

Li Wenjie, Huang Juren, Chen Ying and Jin Peng.2010. Chinese Person Name disambiguation. Http://www.cipsc.org.cn/clp2010/task3_ch.htm.

Lu Yong and Hou Hanqing.2006.Automatic Recognition of Chinese Synonyms Based on Pattern Matching Algorithm. Journal of The China Society For Scientific and Technical Information, 25(6):720-724.

Shi Yingchao, Wang Huizhen, Xiao Tong and Hu Minghan.2011.Personal Name Recognition for Multi-Document Personal Name Disambiguation Task .Journal of Chinese Information Processing, 25(3):17-22.

Sun Jigui, Liu Jie and Zhao Lianyu.2008.Clustering Algorithms Research. Journal of Software, 19(1):53-54.

Yu Shiwen, Duan Huiming, Zhu Xuefeng and Sun Bin.2002. The Basic Processing of Contemporary Chinese Corpus at Peking University SPECIFICATION. Journal of Chinese Information Processing, 16(6):58-64.

Zhang Shunrui and You Hongliang.2010.Chinese People Name Disambiguation by Hierarchical Clustering. Modern library and information technology, 11:64-68.

Zhang Zhufei, Ren Feiliang and Zhu Jinbo.2008.A Comparative Study of Features on CRF-based Chinese Named Entity Recognition. The fourth national conference of information retrieval and content security: 111-117.