Chinese Name Disambiguation Based on Adaptive Clustering with the Attribute Features

Wei Tian, Xiao Pan, Zhengtao Yu, Yantuan Xian, Xiuzhen Yang
School of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, China.

Abstract

To aim at the evaluation task of CLP2012 named entity recognition and disambiguation in Chinese, a Chinese name disambiguation method based on adaptive clustering with the attribute features is proposed. Firstly, 12-dimensional character attribute features is defined, and tagged attribute feature corpus are used to train to obtain the recognition model of attribute features by Conditional Random Fields algorithm, in order to do the attribute recognition of given texts and knowledge bases. Secondly, the training samples are tagged by utilizing the correspondences of the text attribute and answer, and attribute feature weight model is trained based on the maximum entropy model and the weights are acquired. Finally, the fuzzy clustering matrix is achieved by the correlation of Knowledge Base(KB) ID attributes and text attributes for each KB ID, the clustering threshold is selected adaptively based on the F statistic, and clustering texts corresponding to ID are obtained, thus the texts corresponding to each ID are gained followed. For the texts not belong to KB, Out and Other types are obtained by fuzzy clustering to realize name disambiguation. The evaluation result is: P = 0.7424, R = 0.7428, F = 0.7426.

1 Introduction

Person search is an information retrieval way for a specific person, due to the phenomenon of name repetition, therefore, name disambiguation problem becomes more and more important. In recent years, various types of evaluation tasks related to name disambiguation have been launched successively at home and abroad. One task is WPS (Web People Search). WPS is aimed at English names and does not provide any knowledge base, instead it require names referring to the same entity to be clustered together. Another related is the KBP (Knowledge Base Population) task in TAC (Text Analysis Conference) has a named entity disambiguation task, which they use the term entity linking. KBP provides a knowledge base (KB) of named entities. The KB provides a mapping from names to entities. One name can be mapped to many entities. The goal of KBP is to link names occurring in the document to the corresponding entities in KB and to cluster names referring to the same entity, if this entity is not included in the KB. The 2nd task of the CIPS-SIGHAN2012 (CLP2012) [1]—-Named Entity Recognition and Disambiguation in Chinese, can be seen as combination of related tasks in WPS and KBP: First the test names in the document should be judged to be common words or named entities; if a name is predicted as a named entity, participants should further determine which named entity in the KB it refer to; finally, if some names are predicted as named entities that do not occur in the KB, participants should instead cluster these names by the named entities they refer to.

For the name disambiguation, most of the work is concentrated on unsupervised-based or semi-supervised clustering disambiguation method, such as Wang proposed to use the vector space model of web content to do expert evidence-pages clustering disambiguation to solve the multi-document coreference resolution problem to some extent [2]. Bollegala put forward the experts clustering disambiguation solution on key phrases extraction automatically in the context and computing similarity, particularly keyword extraction method depended mainly on the individual information, and the entire extraction process was prone to error cascade phenomenon [3]. Zhou presented a two-stage method for name disam-

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 132–137,
Tianjin, China, 20-21 DEC. 2012

biguation based on exclusive and non-exclusive character attributes, which can improve the disambiguation effect to some extent, but it did not give a clear explanation for threshold selection on the improvement of the hierarchical clustering [4]. Zhang used hierarchical clustering algorithm to solve the multi-text ambiguity issue of Chinese names, though it can better distinguish the names' features, considering verb information as features led to a larger noise introduction without making noise reduction processing [5]. Through the analysis of a large number of name texts, the names' attributes in the text have an important impact on name disambiguation. Therefore, this paper uses the training corpus of the CLP2012 name disambiguation to establish the model of attribute recognition and weight distribution, and applies adaptive clustering method, which can automatically select clustering threshold, to achieve the Chinese name disambiguation.

2 Chinese Name Disambiguation Based on Adaptive Clustering with the Attribute Features

2.1 Corpus Preparation

We must first define the ambiguous name before the name disambiguation. As same as the definition of ambiguous names in CLP2010, the CLP2012 is based on the assumption that "one text corresponds to one person name", that is, supposing a text corresponds to only one person's name, there is no one-to-many problem between the text and the name.

According to text analysis of the CLP2012 training corpus, we found that not all text content is to play a significant role in name disambiguation, but the momentous attributes related to the person appearing in the text is very important to distinguish the persons, for example, there are some sentences about the sports figure inserting into the text of the artistic literature topic, and a few words in above sentences written to the attribute information, such as character's career, just plays the important role in the name disambiguation. Therefore, we need to select the attributes related to the character as the name disambiguation features, which can be named as character attribute features, including 12-dimension, respectively, the person's name (rm), place (dm), organization(jg), career(zy), position(zw), awards (ry), gender (xb), nation(mz), education background(xl), graduate school(byyx), birthday(csrq), works(zp). KB files corresponding to each name must be analyzed to extract ID number and corresponding text messages. And mark the relevant features and do attribute recognition.

2.2 Attribute Recognition

Attribute recognition based on Conditional Random Fields (CRFs) achieves very good recognition effect [6]. Therefore, the CRFs Tools package is used to train on the marked attribute features to obtain the recognition model of 12-dimension attribute feature. The texts and KB files of the CLP2012 test corpus are respectively done the attribute feature recognition by using the recognition model.

Due to the feature of an attribute may be repeated many times, the duplicate must be removed after text recognition completed. Corresponding to each text, there is a feature set $N = \{a_i | i = 1, 2, ..., 12\}$, which$N$represents the text number, $a_i$is the ith dimensional feature. According to the feature dimension defined the feature set will be organized into the form of the feature vector. Similarly, each ID which each xml file of knowledge base contained is corresponding to a set of attribute features. As the 001th text and to the xml ID = 01 of "白雪(Xue Bai)" for examples, the attribute features of specific text and Knowledge Base as shown in Table 1.

2.3 Attribute Feature Weighting

After obtaining the attribute feature vector of text file, use the answer corresponding to the text to mark the answer category which the text belongs to, and then consider the category number as one new feature to add to the attribute feature vector to form a new feature vector, which is regarded as weight training corpus. Then employ the weight training command of the maximum entropy model for training the weights of feature functions on the corpus, namely the attribute weights W_{oi} $(i = 1, 2, ..., 12)$ for the corresponding dimension.

After getting the weight of each attribute feature, the next is matching calculation of the attribute features，that is similarity calculating between the attribute feature set of texts and the KB feature collection on the test corpus. The attribute feature matching problem is considered as the words' matching. The existing matching methods mainly for Chinese words are "HowNet",

Table 1: The representation examples of the attribute feature set and vector of “白雪（Xue Bai）”.

Document Type	Text	KB
attribute feature set representation	001={白雪（Xue Bai），浙江（Zhejiang），浙江代表团（delegation of Zhejiang），歌手（singer），无（null），无（null），无（null），无（null），无（null），无（null），无（null），无（null），无（null）}	01={白雪（Xue Bai），浙江温州（Wenzhou Zhejiang），浙江军区文工团（Military district enter-tainment regi-ment in Zhejiang），歌手（singer），无（null），无（null），无（null），无（null），浙江温州清县小百花越剧团（Xiaobaihua Yueju regiment in Qingxian, Wenzhou Zhejiang），1975年2月28日（birthday），无（null）}
feature vector representation	（001 白雪（Xue Bai）浙江（Zhejiang）浙江代表团（delegation of Zhejiang）歌手（singer）无（null）无（null）无（null）无（null）无（null）无（null）无无（null））	（01白雪（Xue Bai）浙江温州（Wenzhou Zhejiang）浙江军区文工团（Military district entertainment regiment in Zhejiang）歌手（singer）无（null）无（null）无（null）无（null）无（null）浙江温州清县小百花越剧团（Xiaobaihua Yueju regiment in Qingxian, Wenzhou Zhejiang）1975年2月28日无（null））

“Tongyici Cilin” and “Chinese Concept Dictionary” [7]. Word similarity calculated by “Tongyici Cilin” is the closest to the similarity of people's thinking, so Cilin is selected to calculate the similarity. On the basis of analyzing the classification mode and the word-coding table of Cilin [7] and related theory of meanings, the meaning similarity of the two words is calculated according to their meanings coding and the maximum is taken as the similarity finally, the calculating method is shown as follows:

Assume $Sim(x,y)$ is the similarity of the two meanings, if the first letter of the two words' meaning code is the same, then in the same tree$T$, where $T \in \{A,B,C,D,E,F,G,H,I,J,K,L\}$. The formula of $Sim(x,y)$ is shown as follows:

$$Sim\,(x,y) = \begin{cases} f, (x \in T_1, y \in T_2) \\ \delta \times \cos\left(n \times \frac{\pi}{180}\right)\left(\frac{n-k+1}{n}\right) \\ , (x,y \in T_1, C_x \neq C_y) \\ e, \begin{pmatrix} C_x = C_y, \\ C_{xend} = C_{yend} =' \#' \end{pmatrix} \\ 1, \begin{pmatrix} C_x = C_y, \\ C_{xend} = C_{yend} =' =' \end{pmatrix} \end{cases} \tag{1}$$

Where$\delta \in \{a,b,c,d\}$, and if x and y branches at the second layer, then the coefficient$\delta = a = 0.65$, similarly, the third$\delta = b = 0.8$, the fourth $\delta = c = 0.9$, the fifth $\delta = d = 0.96$. In order to control the similarity between 0 and 1, a parameter $\cos\left(n \times \frac{\pi}{180}\right)$ is introduced, where n is the total number of nodes of branch layers, the control parameter $\left(\frac{n-k+1}{n}\right)$ and $k$is the distance between two branches. Define C_x, C_y as the meaning code of x, y, and C_{xend}, C_{yend} is respectively the end symbol of x, y. Take $f = 0.1, e = 0.5$ as a matter of experience.

A large number of statistical results show that two words with the similarity above 0.7 are generally considered to have a similar meaning in people's thinking, so defined that if $Sim\,(x,y) \geq 0.8$, then the attribute features in the same dimension is perceived as matching successful. The weight of the vector matching successful is regarded $W_{oi}*10$ as matching weight $W_{mi}\,(i = 1,2,...,12)$, and the matching weight becomes 0 if the matching is not successful.

2.4 Fuzzy Clustering Matrix Construction

After matching the attribute features between each text and any one ID (short text) in the knowledge

base, the matching weight vector of each text corresponding to the above ID gotten by matching, is the row vector of initial matrix. Since the initial matrix is not square, which is the product of attribute feature matching and not the similarity of the texts in the true sense. All above makes that the adaptive clustering can not work. Therefore transform and adjust on the initial matrix by the cosine of the angle is to make it become the fuzzy clustering matrix of fuzzy clustering.

Assume text set$U = \{x_1, x_2, ..., x_n\}$, where $x_i = \{x_{i1}, x_{i2}, ..., x_{im}\}\,(m = 12)$ is the attribute feature vector, build the fuzzy clustering matrix. The similarity between x_i and x_j is:

$$A = r_{ij} = \frac{\sum_{k=1}^{m} x_{ik} \cdot x_{jk}}{\sqrt{\sum_{k=1}^{m} x_{ik}^2}\sqrt{\sum_{k=1}^{m} x_{jk}^2}} \tag{2}$$

Where x_{ik}, x_{jk} represents the feature vector in the same dimension between texts, and calculate to obtain the similarity matrix, that is fuzzy clustering matrix$A$.

2.5 The Adaptive Clustering Based on the Attribute Features

The Adaptive Algorithm Thought and the Process Description. The fuzzy clustering is a common clustering method in pattern recognition, and has achieved very good effect on pre-classification of characters in Chinese character recognition [8] and classification and matching of speech recognition [9]. Aiming at the task characteristics, different text may has different attribute feature relationships with the knowledge base. If we use the same clustering threshold, it may cause one ID-type clustering better, while another is not good consequences. Thus this paper selects fuzzy clustering method to do name disambiguation processing, according to the difference between the fuzzy clustering matrix generated each time and the content of knowledge base ID, adjust the clustering threshold dynamically and adaptively, and then cluster for each ID of the Knowledge Base through adaptive clustering way. The main idea is to make the classical partition definition fuzzification and dynamically adjust the threshold, which can be solved effectively that 0,1 binary membership can not fully reflect the actual relationship between the data points and the cluster center.

Different thresholds$\lambda \in [0, 1]$can lead to different classifications in Fuzzy clustering analysis, in order to form a dynamic clustering diagram, which makes the classification of the sample image and intuitive. We need to find the optimal λ to effectively cluster some texts with their corresponding ID of KB, and then the clustering result corresponding to λ now is the best result. In this paper, the$F$statistic is used to determine the optimal λ.

Set the text set $U = \{x_1, x_2, \ldots x_n\}$ is the text sample space, and each text$x_i$has $m$features:$x_i = (x_{i1}, x_{i2}, ..., x_{im})\,(i = 1, 2, ..., n)$.Thereby the initial matrix is obtained, where $\overline{x_k} = \frac{1}{n}\sum_{i=1}^{n} x_{ik}\,(k = 1, 2, \ldots, m)$, and $\overline{x}$ represents the center vector of the overall sample, that is any one ID of KB. Set the number of categories is r corresponding to λ , the number of texts is n_j in the jth cluster, $x_1^{(j)}, x_2^{(j)}, ..., x_{n_j}^{(j)}$ is denoted. The cluster center, that is the jth ID of KB, of the$j$th cluster is $\overline{x}^{(j)} = \left(\overline{x}_1^{(j)}, \overline{x}_2^{(j)}, ..., \overline{x}_m^{(j)}\right)$, where$\overline{x}_k^{(j)}$is the average of the$k$th features, namely $\overline{x}_k^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ik}^{(j)}\,(k = 1, 2, ..., m)$.

The F statistic is shown as follow:

$$F = \frac{\sum_{j=1}^{r} n_j \left\|\overline{x}^{(j)} - \overline{x}\right\|^2 \Big/ (r-1)}{\sum_{j=1}^{r}\sum_{i=1}^{n_j} \left\|x_i^{(j)} - \overline{x}^{(j)}\right\|^2 \Big/ (n-r)} \tag{3}$$

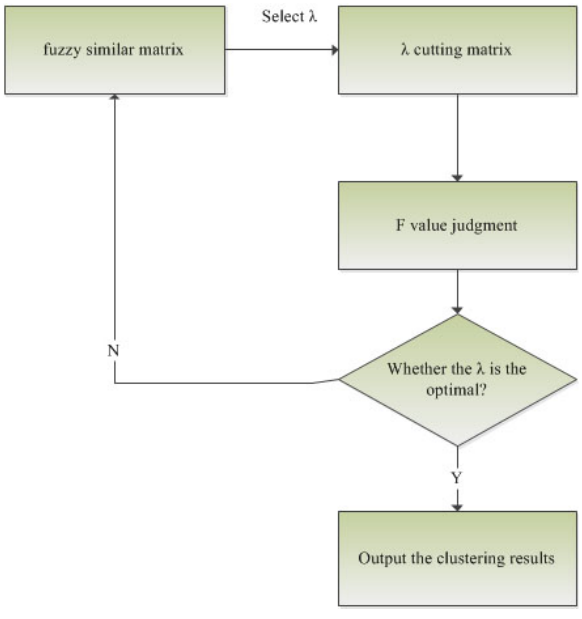

Figure 1: The flowchart of adaptive clustering for name disambiguation

Where $\left\|\overline{x}^{(j)}-\overline{x}\right\| = \sqrt{\sum_{k=1}^{m}\left(\overline{x}_k^{(j)}-\overline{x_k}\right)^2}$ is the distance between $\overline{x}^{(j)}$and $\overline{x}$, $\left\|x_i^{(j)}-\overline{x}^{(j)}\right\|$ shows the distance of $x_i^{(j)}$ and the center $\overline{x}^{(j)}$ in the jth sample. The formula (1) is named F statistic, which follows a Fdistribution with the degree of freedom$r-1, n-r$. For the F statistic, the distance between clusters is represented by the numerator while the distance in one cluster, the denominator. So the larger$F$statistic, the larger distance between clusters, that is the larger distance is inferred between the texts not related to the ID and the texts corresponding to, which shows out a better clustering result.

It can be known that the difference between clusters is significant and illustrates a more reasonable classification result according to the theory of mathematical statistics and analysis of variance, if $F > F_\alpha(r-1, n-r)(\alpha = 0.05)$. If there are more than one F statistics meeting the requirements, $(F-F_\alpha)/F_\alpha$must be examined further, and we can get the maximum of F, which the λcorresponding to is the optimal threshold.

The Realization of the Adaptive Clustering. According to the evaluation task of CLP2012 Chinese name disambiguation, the final answer to the clustering usually consists of three types, namely one is the ID type marked in the "KB", another is the "out" type, which contains not only text attribute features but also not appeared in the Knowledge Base. Besides, there is an "other" type not containing entities and considered as ordinary word. So each type is processed respectively in this article.

For "KB" type, firstly the attribute feature correlation of KB ID and text is used to obtain fuzzy clustering matrix for each KB ID. Secondly adaptive clustering threshold is adaptively selected based on the Fstatistic, and the clustering result corresponding to the threshold is acquired, that is, the texts corresponding to the above threshold. Finally these texts clustered should exclude, and then the rest of the texts and the next ID is used for clustering. Repeat the above process until the rest of the text can not be clustered into a group with the KB ID. For the "other" type, if the texts not related to the KB are extracted no attribute features, and then these texts are regarded as "other" type. For the "out" type, clustering, , a text in the texts excluded the "KB" type and the "other" is randomly selected as a basis for the matching of attribute features, and fuzzy clustering matrix is obtained, then clustering threshold is adaptively chosen to get the clustering result according to the F statistic.

3 Experiments

3.1 Experimental Data

Table 2: Experimental data statement.

Experimental Data	The number of text set	The number of text in each text
training data	16	50-200
test data	32	50-500

There are two types of data given in the evaluation. One is knowledge base, NameKB. A XML file for each test name is provided. This file contains several entries describing the name. The file is named as Name.xml, where Name is the test name. For example, the file for 雷雨(Yu Lei) is 雷雨(Yu Lei).xml. Another is text collection, T for each test name. All texts containing the name N are placed under the folder N. For example, all text containing 雷雨(Yu Lei) are under the folder 雷雨(Yu Lei). Every file in the folder is a plain text file, named as XXX.txt, where XXX is three numbers.

The evaluation tool used in the experiment is provided by the evaluation project group of CLP2012. The overall evaluation indexes are precision, recall and F-value for all test names.

3.2 Experiment Results and Analysis

Do the experiment on the test data by using our approach. The evaluation results are given as follows:

Table 3: The evaluation index comparisons of training data and test data.

DataSets	Precision	Recall	F-value
training data	0.9256	0.9032	0.9143
test data	0.7424	0.7428	0.7426

As can be seen from the results, the attribute recognition model for name disambiguation has

taken good effect. The identification effect is better on training data than the test data. Analyzing the reasons, the recognition errors may be caused by a variety of reasons. For example, the error that the original text is related with name but identified to common word that accounted for 1/2 of the error portion. According to the statistics, the error distribution is shown in the following table.

Table 4: The distribution of the attribute feature recognition errors.

Error Types	Error Proportion
names are recognized to common words	0.5162
only recognized a part of names	0.1956
common words are recognized to names	0.0659
the most important attribute is not identified	0.2223

4 Conclusion

For the characteristics of the evaluation task CLP2012 named entity recognition and disambiguation in Chinese, a Chinese name disambiguation method based on adaptive clustering with the attribute features is proposed, which will resolve this complex disambiguation task into KB type, out and other three types for processing. Do the attribute recognition of given texts and knowledge bases，using the recognition model of attribute features trained by Conditional Random Fields algorithm. Then the attribute feature weight model is trained by utilizing the corresponding attribute feature with answer tag based on the maximum entropy model. After that, the initial matrix is obtained by matching and weighting on the attribute features, on which the fuzzy clustering matrix is generated by transforming, and then clustering by the adaptive method. The algorithm is characterized in automatically finding the optimal clustering threshold to realize name disambiguation according to the different contents of the text and knowledge base. Further research will focus on non-attribute feature selection and the clustering method optimization.

Acknowledgments

This paper is supported by National Nature Science Foundation (No.61175068), and the Open Fund of Software Engineering Key Laboratory of Yunnan Province (No.2011SE14), and the Ministry of Education of Returned Overseas Students to Start Research and Fund Projects. We appreciate the help and assistance of Yu Qin and Wenxu Long.

The corresponding author is Zhengtao Yu(ztyu@hotamil.com).

References

CIPS, SIGHAN. 2012. *http://www.cipsc.org.cn /clp2012/task2.html*. Tianjin, China.

Houfeng Wang, Zheng Mei. 2005. Chinese multi-document person name disambiguation. *High Technology Letters*, 11(3):280–283.

Bollegala D, Matsuo Y, Ishizuka M. 2006. Disambiguation person names on the Web using automatically extracted key phrases. *Proceedings of the 17th European Conference on Artificial Intelligence*, 553–557. Riva del Garda, Italy.

Xiao Zhou, Chao Li, Minghan Hu, Huizhen Wang. 2006. Chinese Name Disambiguation Based on Exclusive Character Attributes. *The 6th CCIR Conference Proceedings*, 2010:333–340. Harbin, China.

Shunrui Zhang, Lianghong You. 2010. Chinese People Name Disambiguation by Hierarchical Clustering. *New Techology of Library and Information Service*, 2010(11):64–68.

John L., Andrew M., Fernando P.. 2001. *The ICML2001 Proceedings*, 2001:282–289.

Jiule Tian, Wei Zhao. 2010. Word Similarity Algorithm Based on Tongyici Cilin in Semantic Web Adaptive Learning System. *Journal of Jilin University (Information Science Edition)*, 2010,28(6):602–608.

Da Lu, Mingpei Xie, Wei Pu. 2000. A Character Preclassification Method Based on Fuzzy Structure Analysis of Typographical Characters. *Journal of Software*, 2000,11(10):1397–1404.

Xiangdong Yu, Xiuyun Suo, Jianren Zhai. 2002. Speech Recognition Based on Fuzzy Clustering. *Fuzzy Systems and Mathematics*, 2002,16(1):75–79.