The Research of Chinese Name Entity Disambiguation Based On Word Sense Disambiguation

Gang Wang

University of Shanghai for Science &Technology

Shanghai 200093, Peoples Republic of China

rogerwg555@gmail.com

Keywords: named entity disambiguation; the similarity of word sense

Abstract. Named Entity Disambiguation (NED) refers to the processing of determining the real entity concept of a given name with some necessary context. This paper introduces a calculation model of the NED based on the word sense disambiguation in order to solve the problem of Chinese name disambiguation which we will face in the text automatic processing module of the software platform called "Shanghai Academic Degree and Postgraduate Education Information Web". The main idea of the model is to calculate the similarity between the all person templates with same name in the person entity library (PEL) and the real person template extracted from the context which is based on the method of word sense disambiguation and finally determine the most possible person after the matching process.

Introduction

The Project Background. "The 12th five-year plan for the national education development" emphasized on developing the education management system, decision support system, monitoring and analysis system and socially oriented education information service system in the part of "speed the implementation of education information strategy". Under this background, developing the software platform called "Shanghai Academic Degree and Postgraduate Education Information Web" has great realistic significance to promote the process. This paper takes solving the problem of NED as the primary task which emerges in the text automatic processing module of the automatic generation system of Chinese doctor & master degree thesis edition for concealed evaluation in the software platform, and this module requires to determine the most really possible person which is marked by the name appeared in the thesis.

The Description of the Task. In order to strengthen the intelligence of the information platform, the text automatic processing module requires to complete the following two functions. Firstly, the system can recognize the Chinese name in the thesis automatically. Secondly, it needs to determine the most really possible person entity which is marked by the Chinese name that has been recognized in the first step.

For example, there is a text in the thesis as the following.

"A researcher from the Chinese academy of sciences called "liuqun" puts forward using the similarity calculation of word sense into the application of the machine translation and also giving a definition of it in the article called "The similarity calculation of word sense based on HowNet".

So task 1 is to recognize automatically that "liuqun" is a Chinese name and task 2 is to determine the "liuqun" here may refer to the person who is a researcher studying the natural language processing (NPL) from Chinese academy of sciences instead of the person who comes from the chemistry department of northeast normal university or other person called "liuqun".

Obviously, task 1 belongs to the typical problem of named entity recognition (NER) [1] and the current domestic research of Chinese name entity recognition has achieved good progress, so we no longer discuss it here. Task 2 belongs to the named entity disambiguation (NED) [2], but it is also different from traditional problem of NED due to the thesis's limitation of professional field and it is also only for Chinese name. In the traditional NED, the person marked by one name maybe involved in all walks of life and different occupations, so that we need to collect massive person information

about all kinds of name in order to create a massive PEL which can support us to solve the problem .Mostly we create it by the technology of hunting and extracting information from web resource such as Wikipedia. However, in our task, the person marked by the name appeared in Chinese degree thesis mostly refers to the experts, scholars, professors in their professional field and those person information often have been saved in the existed database of education experts. So we can create our PEL without using network technology temporary.

Related Work

In recent years, NED has been noticed by the researchers and the related evaluation also has some influence in the international and domestic. The task of name disambiguation evaluation on the web people searching organized by UNED called WePS [3] (Web People Search Evaluation Campaign) is emerging in recent years, it has held three times in the workshop of the ACL (2007), the WWW (2009) and CLEF (2010). The evaluation regards the name disambiguation as the main task so that it reflected its importance in web people search. However, the task of WePS most on the English name, in 2012, CLP2012 [4] carried out the evaluation task on Chinese name for the first time and the name disambiguation in this evaluation had been regarded as a clustering problem.

The existing study on Chinese name disambiguation mostly belong to the method of document clustering and among them the clustering based on Vector Space Model (Vector Space Model, VSM) [5] and social networks [6] is more popular. The clustering based on VSM mostly mainly use the similarity between the context with name in the document to distinguish the documents and the name in documents. And the clustering based on social network use the people and social relation to classify document.

We will adopt another method by calculating the similarity between the all person templates with same name in PEL and the real person template extracted from the context instead of the method of document clustering in this paper.

The Overall Process of Name Disambiguation

As we have no need to concern task 1, in other words, the system has already recognized Chinese name in the thesis automatically. Therefore, we could put a Chinese name directly which we want to disambiguate as the input information of the system. The overall implementation process of name disambiguation is showed in the figure 1.

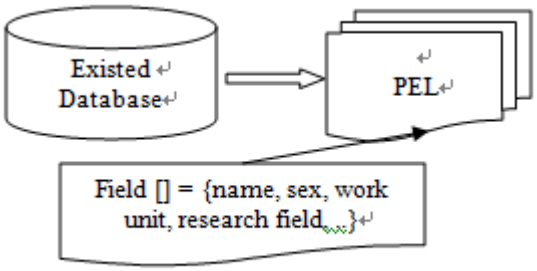

Figure1

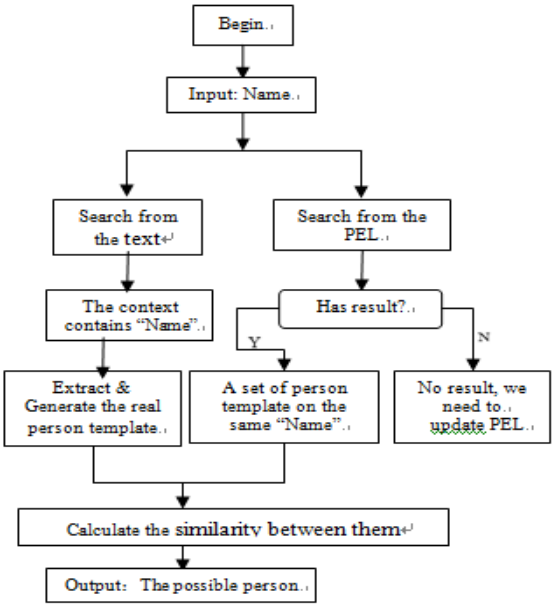

Figure2

From the process, we will see the following three problems need to solve in the disambiguation task: creating the person entity library (PEL), extracting the real person template from the context which contains the name and the matching model based on the similarity calculation.

Creating the person entity library (PEL). Different from the way adopted by the traditional NED which creating the PEL by the technology of hunting and extracting information from web resource such as Wikipedia. We create the PEL by just extracting the most important fields which can represent people's characteristics from the existed database of education scholars into the PEL such as name, sex, work unit, research field and so on. Obviously, the more important field should has a bigger weight value that we will introduce in the later.

Then we can regard the PEL as a mount of sets contains the attribute fields about all kinds of name. In other words, the PEL has saved the existed person templates of the Chinese name. It likes the figure 2 above.

Extracting the real person template from the context. We will extract the same category word set from the context about all fields which is chosen in the PEL and then save it into the new real person template set accordingly. It represents the real person's characteristics in the text. The technology used during the extracting we will describe in another paper.

If some fields in the PEL cannot find the corresponding word with same category in the context, we can add a weight value on the field and set it 0. The process like figure 3.

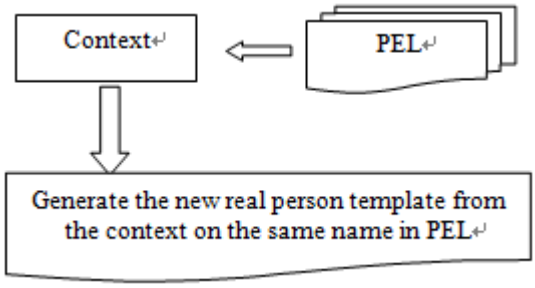

Figure 3

The matching model of name similarity calculation will be described in the next section.

The Similarity Calculation Model of Chinese Name Based On Word Sense Disambiguation

Word Sense Disambiguation. As a nearly research field, word sense disambiguation (WSD) [7] refers to determine the meaning of the word according to the context automatically. It is also a fundamental key research topic in the field of computational linguistics. So it is reasonable for us to analysis the relation between them. As a job of semantic analysis, the core problem of word sense disambiguation is needing a corpus as the source tool of knowledge provided. Here we choose to use HowNet corpus as the tool that determine the semantic relationships.

The core of word sense disambiguation is the semantic similarity calculation. Semantic similarity is an important index of words relationship measurement. Similarity is defined as the degree of that two words in different contexts but can be replaced by each other without changing the text of the syntactic and semantic structure by Liu Qun etc. They also put forward the calculation model of semantic similarity which is based on HowNet [8]. Assuming that two words respectively as W_1 and W_2, their semantic similarity will be like the equation (1).

$$\mathrm{Sim}(W_1, W_2) = \mathrm{Max}\ \{(S_{1i}, S_{2j})\}\ (i=1. .n;\ j=1. .m) \quad (1)$$

S_{1i} is the i^{th} mean of W_1; S_{2i} is the j^{th} mean of W_2

The calculation model is mainly used to calculate the ambiguity of Chinese word segmentation. The value of the "sim" we can get by using the HowNet tool. After calculating the similarity between W_1 and W_2 with different kinds of semantic options, we take the maximum value of the "sim" as the semantic similarity value.

The Similarity Calculation Model of Chinese Name Based on Word Sense Disambigution

Like the calculation model of the Word Sense Disambiguation, we regard a Chinese name as an object waiting for disambiguating instead of a word. We suppose that a name links i person entity templates in the PEL, which means the name has i options for different link connected to different person and each option is marked as Ni. In addition, we suppose one name option has j fields in the name set saved in PEL that can represent the character of each person. We also mark each field here as [Wi1, Wi2,,Wij], then we can regard one person template option in the PEL as a set like the following：N[i] = {Wi1, Wi2,.., Wij}

Then we get the real person template from the context which is also represented by a set of words as S[], we suppose we extract j kinds of word from the context by default, but there is no possible that we could find all corresponding word about every field saved in the PEL. So we add a weight value on the word as following.

S [] ={X1*Y1, X2*Y2,, Xj*Yj} (X1..Xj= {0, 1}) (2)

In the representation (2), Y[j] means the corresponding word to j fields in the PEL extracted from the context. The weight Xj only values 0 or 1, which means there is no corresponding word to some field in the PEL when its value equals 0.

At this step, we successfully represent the name entity not only from the context but also saved in the PEL as a field set. So we can define the similarity of name as the following equation 3.

Sim(S, N) = Max {Sim (S, Ni)} (3)

Then we change the name similarity calculation into word similarity computation like equation 4.

$$\mathrm{Sim}(S, N_i)=\sum_{t=1}^{j} a_t * \mathrm{Sim}\ (S[t],N[t]) \quad (4)$$

Note: The weight at exists because each field has different influence on weighting the value on that how the field can represent the person's feature in degree. The real approximation could be counted after thousands of experiment.

However, there should be a threshold 'm' existed and when the sim(S, N) < m, we should think there is no reliable person entity in the PEL. Only when sim(S, N)>=m, we may consider that the Ni is the most reliable person entity in the PEL and we could output the person entity as the final most possible result at last.

Summary

In this paper, the main job is providing a solution to solve the problem of NED in the text automatic processing module of the software platform called "Shanghai Academic Degree and Postgraduate Education Information Web". And we also arise a name similarity calculation matching model based on word sense disambiguation. The scheme has the practical and application value and it will be used in the development of the software platform after lager amount of train experiment in the later.

However, there are still a lot of work to be done. For example, there is no a large number of train experiment to give the parameter value in the model temporary, but it will be continuously solved in the future. In addition, gathering more sources to expand the person entity library will also be a big job in the future. We will consider real-time grasping information from the network resources, such as Wikipedia and it maybe appear another challenge about stability and security of the system.

References

[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and Classification," Lingvisticae Investigations, vol. 30, no. 1, pp. 3–26, 2007.

[2] S. Cucerzan, "Large-scale named entity disambiguation based on Wikipedia data," in Proceedings of EMNLP-CoNLL, vol. 6, 2007.

[3] Y. Chen, S. Y. M. Lee, and C.-R. Huang. Polyuhk, “A robust information extraction system for web personal names”. In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.

[4] CLP2012.http://www.cipsc.org.cn/clp2012/task2-cn.html

[5] Bagga A, Baldwin B, “Entity-based cross-document conferencing using the vector space model”. In Proceedings of the 17th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1998:79-85.

[6] Langjun, Qinbin and Songwei, “The name disambiguation of the result of name searching based on social network” Chinese Journal of Computers.2009, 32(7):1355-1374

[7] Chan, Y. S., H. T. Ng and D. Chiang, “Word sense disambiguation improves statistical machine translation ”.Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL). Pp.33- 40. 2007.

[8] Liuqun, Li sujian, “The similarity calculation of word sense based on HowNet” The 3rd Chinese word sense workshop, 2002.