Chinese Named Entity Recognition and Disambiguation Based on Wikipedia

Yu Miao, Lv Yajuan, Liu Qun, Su Jinsong, and Xiong Hao

Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, China 100190
{yumiao,lvyajuan,liuqun,sujinsong,xionghao}@ict.ac.cn

Abstract. This paper presents a method for named entity recognition and disambiguation based on Wikipedia. First, we establish Wikipedia database using open source tools named JWPL. Second, we extract the definition term from the first sentence of Wikipedia page and use it as external knowledge in named entity recognition. Finally, we achieve named entity disambiguation using Wikipedia disambiguation pages and contextual information. The experiments show that the use of Wikipedia features can improve the accuracy of named entity recognition.

Keywords: swikipedia, named entity recognition, named entity disambiguation.

1 Introduction

A large of new information emerged and formed information explosion with the rapid development of information technology and Internet. Many emerging information processing technologies such as information retrieval, information extraction, data mining and machine translation appeared in this background. Named entity is the main carrier of information and it expresses the main content of the text. It is the very import part of these researches. The research on named entity recognition has strategic significance to language understanding and information processing.

At present, there are a lot of researches on named entity. The methods can be divided into three types. One is rule-based method. Its effect is good, but writing rules is time-consuming and labor-intensive and it lacks field adaptive capacity. The second is statistics-based method. Although statistics-based method has a good ability of model learning without human intervention, it is limited by the limited scale of the training corpus. As a result the last work emerged which combine rule-based method and statistics-based method. It aimed to reduce the complexity and blindness of the rule-based method. In recent years, a large number of new words are emerging and most of them are named entity including person names, location names and organization names. Traditional rule-based or statistics-based method can't satisfy the named entity recognition and translation tasks, because of the accelerated update speed and expanding scale. In this work, we research on named entity recognition based on network resources in order to improve the performance of the tasks.

M. Zhou et al. (Eds.): NLPCC 2012, CCIS 333, pp. 272–283, 2012.

This article studies basic page, disambiguation page, redirection page, structured categories, hyperlinks and information box. First, we establish Wikipedia database using open source tools named JWPL. Second, we extract the definition term from the first sentence of Wikipedia page and use it as external knowledge in named entity recognition. Finally, we achieve named entity disambiguation using Wikipedia disambiguation pages and contextual information. The experiments show that the use of Wikipedia features can improve the accuracy of named entity recognition.

2 Wikipedia

Wikipedia is a multilingual, web-based, free-content encyclopedia. Since its creation in2001,Wikipedia has grown rapidly into one of the largest online encyclopedia attracting the majority of Internet users to contribute their knowledge to achieve a huge amount of data sharing .There were 441,405 articles in Chinese and 3,917,431 articles in English. Next we will describe basic page, disambiguation page, redirection page, structured categories, hyperlinks and information box of wikipedia in detail.

2.1 Basic Page

A basic page is also known as an entry which describes a real world entity or concept corresponding to a subject. Basic page has a simple title which usually corresponds to the standard name of the entity. Alternative name and abbreviation are defined on redirection page and linked to this page. The first few paragraphs of the basic page especially the first sentence give us the definition and basic description of the entity concept. The following paragraphs expand the detailed description of the entity from all angles on the topic.

2.2 Redirection Page

Entities in the real world usually have two or more names. These different names describing entities of the real world are synonyms. Redirection page in the wikipedia is to solve the synonym problem. Only one of the most representative words in synonyms in Wikipedia is the title of the basic page and the others are titles of redirection pages. When the word matches to the redirection page it will be automatically re-link to the basic page which have the real description of the entity. For example, we entered a word named CAS in the search box. CAS is a redirection page ,so it is directly redirected to the basic page named Chinese Academy of Sciences.

2.3 Disambiguation Page

Disambiguation page is used to deal with the ambiguous name. The so-called ambiguity means the same name may refer to different entities. For example, Washington may refer to President George Washington, it may also refer to Washington State。

Disambiguation page in wikipedia lists the entries which may refer to different entities and contains links to the basic page. It does a brief introduction of every entry in a sentence. For example, we input Washington in the search box. Its disambiguation page lists wikipedia entry named George Washington(first President of the United States of America). Click it and you can enter the corresponding basic page. It also list the entry named Washington State(a state on the Pacific coast of the United States of America) and many other entries related with Washington.

2.4 Categories

Wikipedia provide a grid-like classification system which is edited by the public. An entry belongs to one or more categories in the classification system. A Category is usually constituted by a noun phrase which describe the entity attributes or type information. For example, entity Li Ning belongs to the category ”1963 births”, “living people”, “Chinese male artistic gymnasts”, “Chinese businesspeople” and so on. In addition, each category has its own parent categories and subcategories. For example, category” Chinese business people” has four parent categories such as “Businesspeople by nationality” and “Chinese people by occupation” and six subcategories such as “Chinese real estate businesspeople” and “ Hong Kong business people”. In this way, Wikipedia's classification system constitutes a hierarchical structure which is not a tree in strict sense, but a directed acyclic graph.

3 Named Entity Recognition based on Wikipedia

3.1 Extract Wikipedia features

We do a summary introduction of wikipedia in the second quarter. The first few paragraphs of the basic page especially the first sentence give us the definition and basic description of the entity concept. We can understand the properties of the entity according to the first sentence without reading the full text. The following lists a few examples of the first sentence of the Wikipedia:

（1）The **Chinese Academy of Sciences** (**CAS**), formerly known as **Academia Sinica**, is the national academy for the natural sciences of the People's Republic of China.

（2）**Li Ning** (Simplified Chinese: 李宁; Traditional Chinese: 李寧; Pinyin: Lǐ Níng; born March 10, 1963 in Laibin, Guangxi) is a well-known Chinese gymnast and entrepreneur.

（3）**Beijing** is the capital of the People's Republic of China and one of the most populous cities inthe world

It can be seen from the example, the core term in the noun phrase after the defining verb is a very good knowledge which reflects the attributes of the entry. We extract the core terms to observe,“The Chinese Academy of Sciences is a academy”，“Li Ning is a entrepreneur” , “Beijing is a city” . We can see that this definition term "agency", "entrepreneur", "city" help us to judge an entry is a name, a local name or a organization name. Therefore, we want to extract the core terms used in the first sentence from the Wikipedia article as an additional source of knowledge added to the process of named entity recognition. The specific steps to extract the wikipedia features are as follows:

（1）Do word segmentation and part of speech tagging in the first paragraph of wikipedia

（2）Extract the core term after the defining verb in the first sentence of the first paragraph

（3）Extract the core term in the last sentence of the first paragraph if there is no defining verb

（4）The wikipedia features extracted by the above steps are shown in Table 1

Table 1. Examples of wikipedia features

wikipedia entry	wikipedia feature	wikipedia entry	wikipedia feature
Donald Ervin Knuth	Professor	Li Ning	Entrepreneur
Euskara	Language	Yunnan	Province
Young learn ourselves	Reading book	Zhangguorong	Entertainer
Soft-WorldInternational Corporation	corporation	Wenjiabao	Premier

3.2 Add Wikipedia Features to Named Entity Recognition

We usually annotate corpus using IOB2 tags as we represent named entities. InIOB2 tagging, we use “B-X”, “I-X”, and “O” tags, where “B”, “I”, and “O” means the

Recently	O
Stefanie	B-singer
Sun	I-singer
’s	O
album	O
sold	O
well	O

Fig. 1. the marked results with wikipedia features

beginning of an entity, the inside of an entity, and the outside of entities respectively. Suffix X represents the wikipedia feature of an entity. In the process we apply the knowledge of Wikipedia, the suffix "X" represents the wikipedia features. For example, given a sentence "Recently, Stefanie Sun's album sold well". For example, if we search for "Stefanie Sun" is a Wikipedia entry and extract its wikipedia feature"singer", the marked results of this sentence with wikipedia feature are shown in Figure 1.

4 Experiments and Analysis

4.1 Experiments of Extracting Wikipedia Feature

The test set used in this article are randomly selected from the 372,969 Wikipedia page in wikipedia feature extraction module. It contains a total of 1000 pages. We extract Wikipedia features on the 1000 pages and get the correct rate of 91.5%. Because some sentences are too long, there exists extraction errors .For example, "Information science(or information studies) is an interdisciplinary field primarily concerned with the analysis, collection, classification, manipulation, storage, retrieval and dissemination of information" . For example, "Information science (or information studies) is an interdisciplinary field primarily concerned with the analysis, collection, classification, manipulation, storage, retrieval and dissemination of information." The wikipedia feature extracted by our method is object which is obviously wrong. The correct result should be subject. Because it is relatively long, the final properties fell on the core term of the last clause other than the first clause. First sentence is the defining sentence, it is usually relatively short, entity attribute is in the first sub-sentence in most of the situations. To solve this kind of error, we can do syntax analysis specially on long sentences to find the core words of the parsing results. This article does not introduce a complex syntactic analysis by taking into account the good correct rate of this method coupled with the high cost of parsing.

4.2 Experiments in Named Entities Coverage of Wikipedia

This article use 863 named entity evaluation corpus in 2004 which contain 367 documents and 19,102 sentences. There are 30,955named entities in this 19,102 sentences(named entity in this article specially refers to person name, location name and organization name). LDC dictionary contains a large number of named entities and proper nouns. This article compares wikipediaentry with LDC dictionary. We respectively use them to math 863 named entity evaluation corpus to test the named entity coverage of wikipedia. Test results are shown in Table 2.

As can be seen from Table 4.5, The named entity coverage of Wikipedia is 12.18% higher than that of the LDC dictionary. Although the scale of wikipedia entries are far smaller than the LDC dictionary(788, 745 less),the matched entries with wikipedia have only 4,374 less than that with LDC dictionary in 863 named entity evaluation corpus. So we can see that the named entity coverage of wikipedia is high.

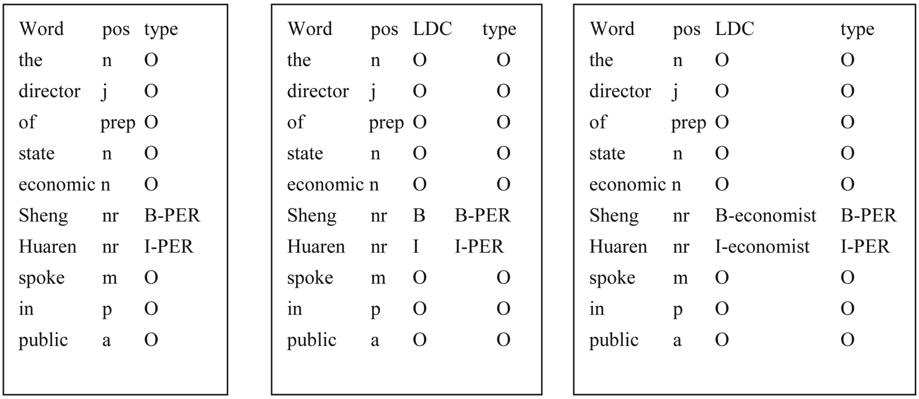
Table 2. Test results in named entity coverage of wikipedia

	The scale of entry	The number of matched entries	The number of matched NE	NEproportion
LDC dictionary	1,161,714	32,611	13,151	53.74%
Wikipedia	372,969	19,951	17,525	65.92%

4.3 Experiments of Named Entity Recognition Based on Wikipedia

In this section, we demonstrate the usefulness of the wikipedia feature for NER. We divide the 863 named entity evaluation corpus into two sets. One is the training set including 17, 102 sentences and the other is the test set including 2000 sentences. We use CRF as the classifier of named entity recognition and do three tests. We carry out the first test with the common features such as word and POS, the second test with LDC dictionary in additional to the common features and the third test with wikipedia feature in additional to the common features.

We annotate the 863 named entity evaluation corpus in a word sequence. “Sheng Huaren” in the example have appeared both in the LDC dictionary and wikipedia and its wikipedia feature is “economist”. The annotated corpus is shown in Figure 2.



Word	pos	type
the	n	O
director	j	O
of	prep	O
state	n	O
economic	n	O
Sheng	nr	B-PER
Huaren	nr	I-PER
spoke	m	O
in	p	O
public	a	O

common features

Word	pos	LDC	type
the	n	O	O
director	j	O	O
of	prep	O	O
state	n	O	O
economic	n	O	O
Sheng	nr	B	B-PER
Huaren	nr	I	I-PER
spoke	m	O	O
in	p	O	O
public	a	O	O

LDC feature and common features

Word	pos	LDC	type
the	n	O	O
director	j	O	O
of	prep	O	O
state	n	O	O
economic	n	O	O
Sheng	nr	B-economist	B-PER
Huaren	nr	I-economist	I-PER
spoke	m	O	O
in	p	O	O
public	a	O	O

wikipedia feature and common features

Fig. 2. Annotated 863 named entity corpus

This article uses CRF++ tool, the above defined three feature templates and corresponding three annotated corpus to test the performance of the named entity recognition. The test results are shown in Table 3 and 4 :

Table 3. Test results of named entity recognition（in word sequence）

	P	R	F
Common features	83.13	87.57	85.35
LDC feature	83.52	87.84	85.68
Wiki feature	84.46	88.81	86.64

Table 4. Test results of named entity recognition（in character sequence）

	P	R	F
Common features	82.24	84.89	83.56
LDC feature	82.51	85.27	83.89
Wiki feature	82.77	85.75	84.26

As is seen from experimental results, the Wikipedia features improved the accuracy in F-measure by 1.29 points(in word sequence) and 0.7 points(in character sequence) compared with common features. Compared with LDC dictionary, the Wikipedia features improved the accuracy in F-measure by 0.96 points and 0.37 points respectively. The wikipedia feature can play a good role in named entity recognition.

The simple method that extracting defining feature from the Wikipedia page can effectively improve the correct rate of named entity recognition. The results show that the structured features of Wikipedia is conducive to extraction of knowledge. Our method is simple but effective because of the following reasons: (1) If a Wikipedia page is a disambiguation page, we will not extract the defining feature of its corresponding Wikipedia, so we will not bring the wrong noise where it might be existed. (2) If a Wikipedia page is non-ambiguous pages, it will describe the main meaning of most editors agreeing on this entity. The reason that the extracting Wikipedia definition features is helped to improve the correct rate of named entity recognition is the main meaning of the entities are frequently used in the corpus. In our method, there is still room for improving , for example, the disambiguation of ambiguous entities. If the Wikipedia pages make continuous rapid growth at the current rate, perhaps all of the Wikipedia entity will become ambiguous entity in the latest future. We need a disambiguation method to find the most appropriate page from the multiple pages listed in the disambiguation pages.

5 Named Entity Disambiguation Based on Wikipedia

Entity ambiguities refers to an entity alleged corresponding to the problem of real-world entities. For example , the following three entities alleged "Washington" ：

U.S. founding fathers of Washington ；

Washington, DC, the capital of the United States.

Washington, located in the northwestern United States.

They separately refer to the three real-world entity ： "America's first president"、"the U.S. capital " and "the U.S. state of Washington". In the task of Named entity recognition, the alleged "Washington" may be a person name (George Washington), or a local name (Washington, DC, or Washington), and we need to determine which the true type of the alleged entity are belonged to , and this is named entity disambiguation.

5.1 The Processes of Named Entity Disambiguation Based on Wikipedia

Wikipedia disambiguation page lists the entity alleged ambiguities. And it provides us with a good disambiguation information. If there is "Washington" in a sentence, and "Washington" is an ambiguous entity, it may be a person name or a local name. The original model of random condition is not very accurate in classifying. And the "American President" entry which "George Washington" entry at, where the Wikipedia disambiguation page list , can help us determine to decide "George Washington" is a name. Besides another cited "Washington State" entry where it has the "American states" entry, can also let us be certain about the "Washington State" is the local name. Through the Classification and Labeling of Wikipedia entries we can properly mark the category of the named entities they belong to in 80% of Wikipedia coverage rate、95% of correct rate. Intuitively, we can determine the "Washington" should be which entity through the current context of a sentence, so we think about a method : we calculate the similarity of all the pages listed in the sentence of the documentation and Wikipedia disambiguation page, then find out the most similar Wikipeia page to the current sentence, finally give a more accurate label of the current entity through the entity identified by the Wikipedia. Therefore, we propose to build the double-layer CRF for named entity disambiguation based on context information , the specific process is shown in Figure 3:

5.2 Training Wikipedia Corpus with CRF

All of the existing Wikipedia entries in Wikipedia page are marked with the symbols of "[[]]", so we use the classification label ,from Wikipedia named entity dictionary (including names dictionary, gazetteer and agencies name a total of 135,504 data dictionary) , to match the entry in the "[[]]" , if the entry exists in the named entity dictionary, we marked the entry with the type of the named entity , otherwise the entry does not exist in the named entity dictionary, the symbol of "[[]]" will be stripped, and ultimately we will convert the Wikipedia corpus with a training data of named entity label. The original Wikipedia corpus and the named entity annotation corpus after converting are shown in Figure 4 and 5, in which person name, place name, organization names marked separately with PER、LOC、ORG .

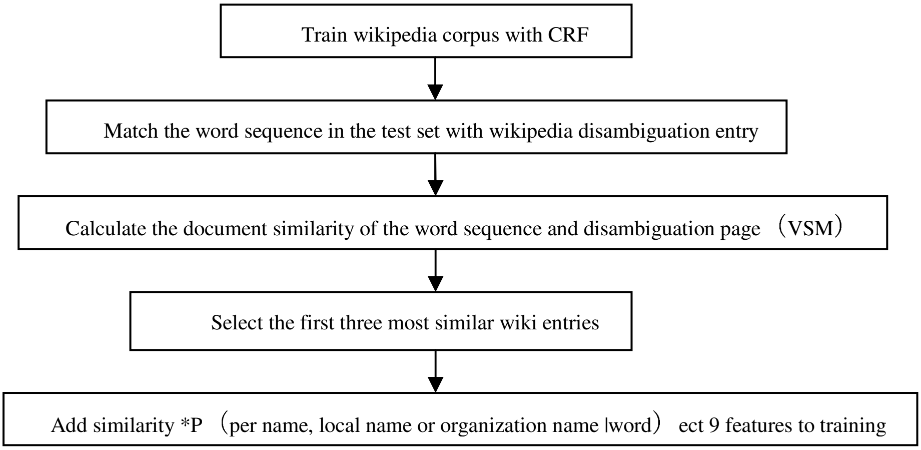

Train wikipedia corpus with CRF
Match the word sequence in the test set with wikipedia disambiguation entry
Calculate the document similarity of the word sequence and disambiguation page （VSM）
Select the first three most similar wiki entries
Add similarity *P （per name, local name or organization name |word） ect 9 features to training

Fig. 3. Named entity disambiguation processes based on Wikipedia

[[Steven Paul "Steve" Jobs]](/ˈdʒɒbz/; February 24, 1955 – October 5, 2011)[6][7] was an American businessman and technology visionary. He is best known as the co-founder, chairman, and chief executive officer of [[Apple Inc.]]

Fig. 4. The original corpus of Wikipedia

[[Steven Paul "Steve" Jobs]]PER (/ˈdʒɒbz/; February 24, 1955 – October 5, 2011)[6][7] was an American businessman and technology visionary. He is best known as the co-founder, chairman, and chief executive officer of [[Apple Inc.]]ORG

Fig. 5. Wikipedia named entity annotation corpus

The correct marked rate of Wikipedia marked corpus named entities after converting is about 95% ,and the recall rate is about 80%. CRF are maked use of to go on training and self re-marking on the label corpus to achieve a higher recall rate.

5.3 The Examples of the Disambiguation of Wikipedia

A simple example can show us the process of the disambiguation. For example, there is a sentence which is "Bloomberg flew to Washington to promote his own ideas" in our testing collection ,then we find the word "washington" is a disambiguaion page on wikipedia through the maximum matching wikipedia entry, and in this disambiguation page lists all the possible ambiguity entry, a total of seven, such as "Washington. DC", "Washington State", "Denzel Washington", these entries have been marked as entity type in the first step a CRF-based training and we have retained the probability of each entity type. We calculate the similarity of the document of the given sentence and the nine Wikipedia page, then find the top three highest similarity of the wiki page which are "Washington State", "Washington DC" and "Washington Town" ,and add nine characteristics of the Similarity (Washington State) * P (person's name | Washington State), similarity (Washington State) * P (local names | Washington State), similarity (Washington State) * P (agency name | Washington State). . similarity ("Washington Township") * P (organization name | Washington town) and so on to the CRF training and testing. It is more intuitive to see Figure 6.

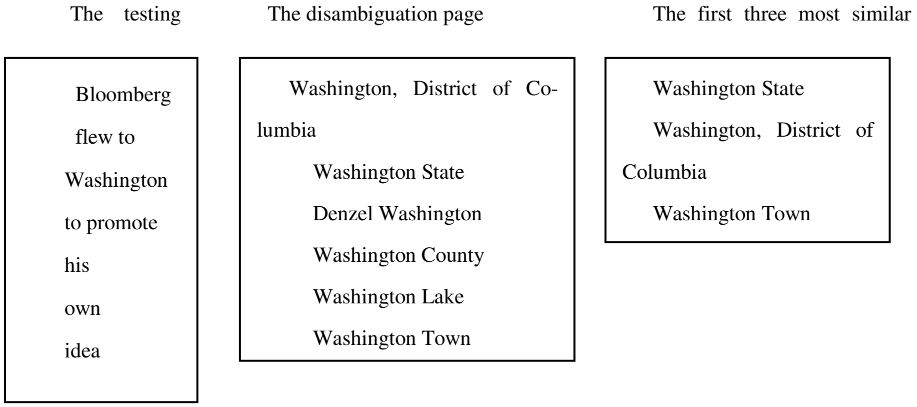


Fig. 6. Example of named entity recognition based wikipedia

6 The Analysis of Named Entity Disambiguation Based on Wikipedia

The corpus used in this study is the same as the terms in section 4th, and we add 9 disambiguation characteristics on the basis of the original wiki features, and the results of the study are shown in Table 5.

The F value is improved by 0.43 in the experimental results, after combining the Wikipedia disambiguation feature,and the effect is significant opposed to the limited corpus number of ambiguous entities .The example of correcting based on the Disambiguation method: "The ITTF has announced the latest world rankings, and the men's singles aspects of German Boer continued in the first place", there are ambiguities of " Boer " in this sentence, "boer" can be place names (refer to South Sudan city)

Table 5. Experimenal results of named entity disambiguation based on wikipedia

	P	R	F
wikipedia feature	84.46	88.81	86.64
named entity disambiguation	84.73	89.42	87.07

or names (refer to table tennis players of Germany).At first "boer" is mislabeled as local name, after combining the Wikipedia disambiguation features, some contextual fea tures, such as Germany, table tennis, men's singles " are used to help the system correcting the category into the right type person name.

7 Summary and Future Work

Firstly we introduces of the defining feature of Wikipedia as an additional knowledge added to the based named entity recognition in the CRF, and the method is simple but effective in improving the rate of correcting th named entity recognition. Then the Wikipedia corpus converted into a named entity annotation corpus is illustrated, this huge label corpus helps us to further improve the performance of named entity recognition, and we make full use of Wikipedia's disambiguation pages and context information to build a double-layer CRF system for carrying out named entity disambiguation. The study shows that the method proposed can improve the correct rate of named entity disambiguation.

In named entity recognition, we only make use of the first sentence of the Wikipedia page, so we can consider that making full use of the Wikipedia category labels, hyperlinks, and other rich resources to further improve the correct rate of named entity recognition.

References

[1] Shun, Z., Wang, H.: Named Entity Recognition Research. Modern Library and Information Technology (6), 42–47 (2010)

[2] Zhou, K.: Rule-based named entity recognition. Hefei University of Technology, Anhui (2010)

[3] Li, J., Wang, D., Wang, X.: Chinese organization name recognition based on template matching. Information Technology (6), 97–99 (2008)

[4] Huang, D., Yue, G., Yang, Y.: Chinese local name recognition based on statistics. Journal of Chinese Information 17(2), 36–41 (2003)

[5] Huang, D., Yang, Y., et al.: Identification of Chinese Name Based on Statistics. Journal of Chinese Information Processing (2001)

[6] Wan, R.: Chinese organization name recognition. Dalian University of Technology, Liaoning (2008)

[7] Qiao, Y.: Chinese named entity recognition with the combination of rules and statistics. Shandong University, Shandong (2007)

[8] Kazamaand, J., Torisawa, K.: Exploiting Wikipedia as external knowledge for named entity recognition. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 698–707 (2007)
[9] Nothman, J., Curran, J.R., Murphy, T.: Transforming Wikipedia into named entity training data. In: Proceedings of the Australasian Language Technology Association Workshop, pp. 124–132 (2011)
[10] Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 708–716 (2007)