JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 8, No. 3, September 2014, pp. 137-148

Classifying Articles in Chinese Wikipedia with Fine-Grained Named Entity Types

Jie Zhou*, Bicheng Li, and Yongwang Tang
Zhengzhou Information Science and Technology Institute, Zhengzhou, China
zhoujie.nlp@gmail.com, lbclm@gmail.com, tangyongwang@gmail.com

Abstract
Named entity classification of Wikipedia articles is a fundamental research area that can be used to automatically build large-scale corpora of named entity recognition or to support other entity processing, such as entity linking, as auxiliary tasks. This paper describes a method of classifying named entities in Chinese Wikipedia with fine-grained types. We considered multi-faceted information in Chinese Wikipedia to construct four feature sets, designed different feature selection methods for each feature, and fused different features with a vector space using different strategies. Experimental results show that the explored feature sets and their combination can effectively improve the performance of named entity classification.

Category: Human computing

Keywords: Named entity classification; Chinese Wikipedia; Fine-grained; Feature selection; NER corpora

I. INTRODUCTION

As the largest online collaborative multilingual encyclopedia, Wikipedia contains millions of available articles in 285 languages. Although written collaboratively by volunteers, Wikipedia has been found to have similar coverage and accuracy to Encyclopedia Britannica, in which articles were contributed by experts [1].

With the development of the Knowledge Base Population track at Text Analysis Conference (TAC), an increasing number of scholars have recently focused on the combination of Wikipedia and named entities (NEs). Compared with other resources, Wikipedia contains a large amount of NEs that have a normalized structure and many types of information. These characteristics make Wikipedia an excellent resource for numerous NE applications, such as named entity recognition (NER) [2], entity linking [3], and entity knowledge base [4].

For pre-defined NE types, the manual process of collecting and annotating NEs is a time-consuming task, and requires significant skill. To address this problem, a number of researches have presented methods to classify articles in Wikipedia with NE types, which is usually considered an NE classification problem of Wikipedia articles. By using the result of NE classification, NE gazetteers can automatically be constructed to improve the performance of NER [5], or increase NE type information for a disambiguation system [6]. Another important application is automatic NER corpora generation from Wikipedia [7, 8], which can be used to annotate the NE types of outgoing links, in sentences within Wikipedia articles.

Different languages have different conventions and resources. For the task of English NER, one of the most effective orthographic features is capitalization in English,

Open Access http://dx.doi.org/10.5626/JCSE.2014.8.3.137 http://jcse.kiise.org

Received 1 November 2013; **Revised** 14 June 2014; **Accepted** July 23 2014
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

which aids in generalization to new text of different genres in an NER system. The capitalization feature of aliases in incoming links is also utilized, to identify large portions of non-entity articles in English Wikipedia [9]. However, this feature is nonexistent in some languages, such as Chinese and Arabic, and the diversity of non-entity articles will undoubtedly lead to greater difficulty in NE classification. To improve the performance of NE classification in special languages, some language-dependent features can be utilized. For example, in the Chinese NER system, the NEs of person (PER) type usually start with the common Chinese family names [10] (e.g., '王' [*Wang*], '赵' [*Zhao*]), and the NEs of location (LOC) and organization (ORG) may end with some character tail hints [11] (e.g., '市' [*City*], '局' [*Bureau*]). These characteristics are also effective features for NE classification.

This paper attempts to classify articles in Chinese Wikipedia with fine-grained NE types. The contribution of this paper lies in considering multi-faceted information in Wikipedia, such as some traditional features and novel features (e.g., co-occurrence relation and article title feature), to construct feature sets; in analyzing the characteristics of each feature and designing different feature selection methods; and in fusing different features with a vector space through different strategies, such as increasing weight, or adding new feature terms. To the best of our knowledge, this study is the first to perform such NE classification in Chinese Wikipedia.

This paper is organized as follows: Section II introduces the related work. Section III describes the NE types used in our experiments. Section IV discusses the feature sets extracted from articles in Chinese Wikipedia. Section V describes the datasets and presents the results of the experiments. Section VI concludes the paper.

II. RELATED WORK

The works to classify Wikipedia articles with NE types mainly fall into two categories: *heuristic rule based* and *machine learning based*. Toral and Munoz [12] utilized heuristic rules that consider the number of nouns in the article definition (first sentence of Wikipedia article) to determine NE type. The work of Toral and Munoz [12] is recognized as the earliest research on NE classification of Wikipedia articles. The following methods based on heuristic rules have received considerable attention on Wikipedia category titles; Richman and Schone [13] produced a set of key phrases from English category titles to classify Wikipedia articles with NE types. Nothman et al. [8] created a set of 141 case-sensitive keywords or phrases that matched English category titles.

Methods based on heuristic rules usually achieve high precision. Heuristic rules can be established more easily than those of other natural language processing (NLP) applications, such as NER. However, the problem with this method is limited coverage. Moreover, a large number of Wikipedia articles cannot match any rule. Thus, the methods based on machine learning are generally applicable to all articles in Wikipedia and can be applied independently [14] or combined with heuristic rules [8].

Researches on machine learning mainly focused on the features extracted from Wikipedia articles, and the classifiers adopted to classify the articles. This method is represented by the works of Watanabe et al. [15], who utilized conditional random fields over a graph constructed from the outgoing links in the articles, and classified Wikipedia outgoing links with NE types. Dakka and Cucerzan [14] trained SVM and Naïve Bayes classifiers by using page-based and context features, and their experimental results showed that structural features (such as the data in the tables) are distinctive in identifying the NE type of Wikipedia articles. Saleh et al. [16] extracted features from abstracts, infobox, category, and persondata structure, and improved the recall of different NE types, by using beta-gamma threshold adjustment. Tkatchenko et al. [17] adopted similar features to Tardif et al. [18]. They added a 'List of' feature to the bag-of-words (BOW) representation, and added a boolean feature, which is the result of a binary rule, to increase separability between the articles of NEs and non-entities.

Most previous researches have focused on the English language, which is characterized by a distinction in the capitalization of common and proper nouns. However, for some other languages (e.g., Chinese, Arabic), the number of potential NEs is significantly larger than that for English, because no distinction can be utilized to remove non-entity articles. To make the most of the distinction in the English language, a reasonable assumption is usually adopted that the article in non-English Wikipedia can be classified into the same NE type as the corresponding English article that is associated with cross-language links [7, 13]. However, this method suffers from limited coverage.

Some language-dependent features have also been introduced to improve performance for special languages. Higashinaka et al. [19] devised some features (e.g., article title features) that are related to the Japanese language, and created an extended gazetteer of 200 NE types. Alotaibi and Lee [20] defined language-dependent and independent features in Arabic Wikipedia.

Wikipedia contains a wealth of multi-faceted information, but only part of the available information is used in related researches. More attempts need be made to exploit the information in Wikipedia as intensively as possible. Moreover, the morphological and orthographic features in special languages can also be used to significantly improve performance.

III. NAMED ENTITY TYPE

Wikipedia provides a categorization hierarchy called

folksonomy, which refers to dynamic knowledge organization systems created by communities of distributed volunteers. However, its complicated hierarchies and large-scale categories make it unsuitable for many semantic applications. With the closer combination of Wikipedia and NE technologies, more researches have focused on NE types for their exact definition and wide application.

Different sets of NE types can be adopted for different applications. To determine the set of NE types in this paper, we extend three universal NE types (PER, ORG, and LOC) into fine-grained types. As in the ACE (Automatic Content Extraction) evaluation (cf. ACE 2005), geo-political entity (GPE) includes inhabited locations with a government, such as cities and countries. Facility refers to a functional, primarily man-made structure. For the definition of the three universal NE types, we also principally refer to the ACE evaluation, and give some examples in Wikipedia.

Person Entities (PER): Person entities are not limited to specific persons in real life (e.g., '*Isaac Newton*', '*Bill Gates*'). In fact, such entities also include fictional characters in the work of art (e.g., '*Harry Potter*', '*Sun Wukong*'), and mythological figures (e.g., '*Poseidon*', '*Jade Emperor*').

Organization Entities (ORG): Organization entities are usually corporations, agencies, and other groups of people defined by an established organizational structure. Generally, such entities are classified with the following subtypes: government (e.g., '*Legislative Yuan*'), commercial (e.g., '*Microsoft*'), educational (e.g., '*Peking University*'), non-governmental (e.g., '*People's Power Party*'), etc.

Location Entities (LOC): Location entities are usually geographical entities, such as geographical areas and landmasses, bodies of water, and geological formations. These entities include water-body (e.g., '*Arabian Sea*' and '*Lake Baikal*'), land-region-natural (e.g., '*the Alps*' and '*the Falkland Islands*'), and address. We also add GPE type as a subtype of LOC.

The coverage of each NE type in Chinese Wikipedia is also considered in the process of designing fine-grained types. Finally, we select five coarse-grained types (PER, ORG, LOC, other miscellaneous entity [MISC], and non-entity [NON]), and 18 fine-grained types. Two auxiliary types ('List of …' page and Disambiguation page) are identified by simple heuristic rules. All fine-grained types and their examples are listed in Table 1.

Table 1. Fine-grained NE types and the examples

NE type	ID	Fine-grained NE type	Example
Person (PER)	1	People	Lu Xun (鲁迅)
	2	Non-People	Zeus (宙斯)
Organization (ORG)	3	Government	State Oceanic Administration (国家海洋局)
	4	Company	Nokia (诺基亚)
	5	Education	Tsinghua University (清华大学)
	6	Media	Xinhua News Agency (新华社)
	7	Other	The Beatles (披头士乐队)
Location (LOC)	8	GPE	China (中国)
	9	Celestial	Moon (月球)
	10	Water Body	The Yellow River (黄河)
	11	Other	The Alps (阿尔卑斯山)
Other (MISC)	12	Facility	The Palace Museum (故宫)
	13	Weapon	Sniper rifle (狙击步枪)
	14	Work of Art	Harry Potter (哈利 • 波特)
	15	Period	Jurassic (侏罗纪)
	16	Language	English (英语)
	17	Game	World of Warcraft (魔兽世界)
Non-entity (NON)	18	NON	Suspect (猜想)
'List of …' page (LIST)	19	LIST	List of Automobile Brands (汽车品牌列表)
Disambiguation page (DAB)	20	DAB	Dragon (Disambiguation) (龙 (消歧义))

NE: named entity, GPE: geo-political entity.

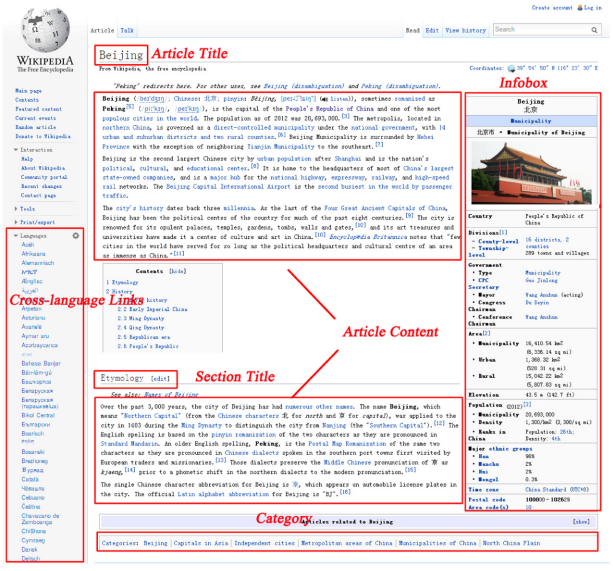


Fig. 1. Information areas in the Wikipedia article '*Beijing*'.

IV. FEATURES EXTRACTION

For the articles in Wikipedia, we prioritize the extraction of features with great discrimination for NE types. Wikipedia editors tend to use the same structure, such as similar section title, infobox, and category, as similar existing articles. However, a large number of articles do not possess any structure information, or possess only some items. For these articles, we should utilize some common features, such as article content and article title, to identify the NE types.

We explore four feature sets, of article content feature, structured feature, category feature, and article title feature. As a common feature set, the BOW representation based on article content is used to build a basic feature vector. Other features can be utilized to optimize the weight of terms, or to build a new feature vector that can be combined with a basic feature vector to classify Wikipedia articles. We consider multi-faceted information in Wikipedia to build our feature sets. Fig. 1 shows an example of information areas in the Wikipedia article '*Beijing*'.

A. Article Content Feature

Article content refers to a detailed description of a Wikipedia article that introduces the related knowledge in textual format. The BOW representation based on article content has been adopted in many researches [14, 19], and is considered a stable feature, because almost all Wikipedia articles contain such information.

Unlike English text, no marked word boundaries exist in Chinese text. We use toolkit NLPIR (http://ictclas.nlpir.org) to realize Chinese word segmentation and part-of-speech tagging, and only keep the terms of the most representative parts-of-speech (noun, verb, and adjective). The feature vector that contains the BOW representation of reserved terms is then built, and the feature weight is computed using the TF-IDF method, as follows:

$$w(t,\vec{d}) = \frac{\log(1+tf(t,\vec{d}))\times\log(N/df_t)}{\sqrt{\sum_{t\in\vec{d}}[\log(1+tf(t,\vec{d}))\times\log(N/df_t)]^2}}$$

where $tf(t,\vec{d})$ is the number of times that term t occurs in document $\vec{d}$, N is the total number of documents in the corpus, and df_t is the number of documents where the term t appears.

Although the feature space is effectively limited by part-of-speech selection, the dimensional space remains large, and the data could be plagued by a substantial amount of noise terms. To reduce feature dimensions, the feature selection method of information gain (IG) is introduced. IG is defined as follows:

$$\mathrm{IG}(t_k) = p(t_k)\sum_{i=1}^{m}p(c_i|t_k)\log\, p(c_i|t_k)$$
$$+p(\bar{t}_k)\sum_{i=1}^{m}p(c_i|\bar{t}_k)\log\, p(c_i|\bar{t}_k)-\sum_{i=1}^{m}p(c_i)\log\, p(c_i)$$

where $p(c_i)$ is the probability of NE type c_i, $p(t_k)$ is the probability that feature t_k occurs in the training set, $p(c_i|t_k)$ denotes the conditional probability of the type c_i that feature t_k occurs, and $p(c_i|\bar{t}_k)$ denotes the conditional probability of the type c_i that feature t_k does not occur.

B. Structured Feature

The articles in Wikipedia are edited by using wiki markup language, and these semi-structured texts are converted into HTML pages by wiki tool. With wiki markup, some important information can be tagged (e.g., section title and infobox template). We discarded outgoing links, because they are more inclined to explain the topics of related articles. Finally, three representative structured features were chosen, namely, section title, infobox, and co-occurrence relation in the lists and tables.

Section Title: Section titles are usually the framework of an article. Words in section titles can indicate NE types more clearly. For example, section titles '*family and childhood*' and '*marriage and children*' are all great indicators of PER type. To combine with the feature vector of article content, we increase the weight of words contained in the section titles, except for some common words, such as '*references*' and '*links*'.

Infobox: Infobox templates are a set of subject-attributes-values triples that often contain a condensed set of important facts relating to the article. The attributes of

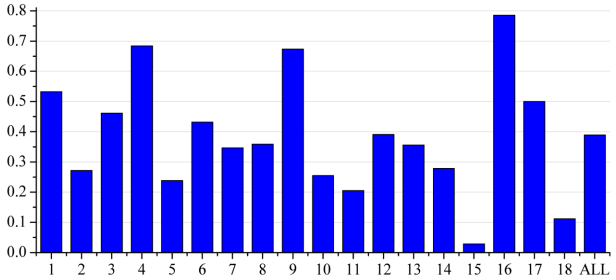


Fig. 2. Percentage of articles that contain infobox template for 18 fine-grained named entity (NE) types. The x-axis represents the ID of fine-grained NE types, and the y-axis represents the percentage. 'ALL' in x-axis is the average percentages of all annotated articles.

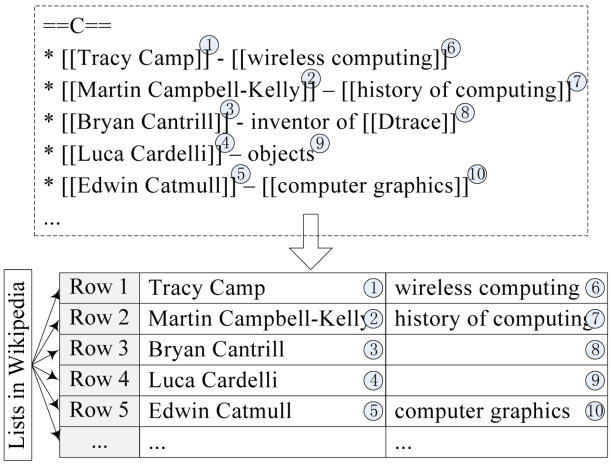


Fig. 3. Transformation of a list into a co-occurrence relation in the *'List of computer scientists'* page. The target articles to co-occurrence articles that occurred in the same column are identified.

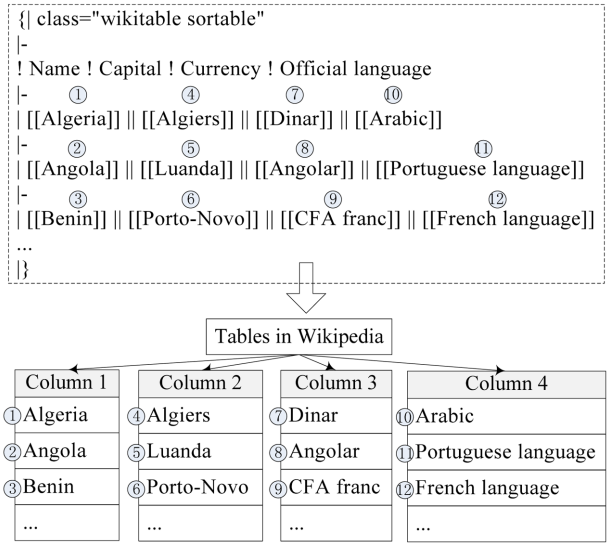


Fig. 4. Transformation of a table into a co-occurrence relation in the *'List of countries and capitals with currency and language'* page. The target articles to co-occurrence articles that occurred in the same column are identified.

infobox are predefined by the infobox template. For example, the infobox '*Infobox Weapon*' lists the attributes '*name*', '*image*', and '*origin*'. Words contained in the values of infobox are processed using the same method as section titles. Approximately 1,300 infobox templates (http://zh.wikipedia.org/wiki/Category:%E4%BF%A1%E6%81%AF%E6%A1%86%E6%A8%A1%E6%9D%BF) involved in different domains are extracted. These templates include '*Infobox Company*' and '*Infobox Automobile*'. As seen in the examples above, the subjects of infobox templates have great discrimination for NE types.

An analysis of distribution of infobox subjects reveals two main problems. First, many Wikipedia articles do not use infobox templates. Fig. 2 shows the percentage of articles that contain an infobox template in the annotated articles of each NE type. Different percentages are found for each NE type. Second, the overlapping subjects of infoboxes could limit the performance. For example, the template '*Infobox Korean name*' could be used in multiple NE types.

Following the analysis above, we build another feature vector based on infobox subject, and add the label 'Infobox' for all features, such as '*Infobox: Infobox Company*'. The weights are computed simply by examining the presence of the subject.

Co-occurrence Relation: The co-occurrence articles, which are tagged with special wiki markup, usually have strong semantic correlation, and the same NE type. Nadeau et al. [21] showed the possibility of creating a list of entities from HTML lists and tables on the Web. Watanabe et al. [15] introduced a graph structure based on the co-occurrence of elements to improve the classification performance of NE types.

Considering that a large amount of noisy results could be identified by complicated structures, we only extracted the co-occurrence relation from the lists and tables in the 'List of …' pages. These hierarchical layouts, especially list items, are usually used to extract hyponymy-relation candidates and to then realize large-scale hyponymy-relation acquisition [22, 23]. Tkatchenko et al. [17] added the 'List of' feature, which is constructed by tokenizing and stemming the titles of all 'List of …' articles containing the current article, to the BOW representation. In this paper, the co-occurrence articles are utilized to extract representative words and increase their weights because only a small number of articles tagged with co-occurrence relation can be found in Chinese Wikipedia. Fig. 3 shows the transformation of a list tagged with wiki markup into a co-occurrence relation of articles in the same NE type. Some lists sometimes have more than one article in each row (Fig. 3). To resolve this issue, we

choose only the same hierarchical articles with the same pattern, such as articles ①, ⑥ and ②, ⑦ in Fig. 3. Similar to Fig. 3, Fig. 4 shows the transformation of a table into a co-occurrence relation. We extract the co-occurrence relation from the same column in the table.

Compared with the total number of Wikipedia articles, the number of articles tagged with the co-occurrence relation remains small (approximately 7%), and the percentage will decrease rapidly, if the article pair is limited to the training and test dataset, respectively. To highlight the commonness expressed by these articles, we extract common words and increase their weights. The detailed process is as follows:

1) For the test article a_i, all co-occurrence articles $\{a_{i1}, a_{i2}, ..., a_{in}\}$ are collected, and the words are extracted from section titles and attribute values of the infobox template in each article a_{ij} (j = 1, 2, ..., n);

2) Only the words that occur in at least half of the co-occurrence articles are reserved;

3) The reserved words and their frequency are added to the feature vector based on the article content of the test article a_i.

C. Category Feature

Categories usually express a relation that is common to all articles in the category, and one article can simultaneously belong to multiple categories that describe the article from different perspectives. For example, the article '*Martin Luther King*' belongs to categories '*Born in 1929*', '*Community organizing*', etc. Some Wikipedia categories are great indicators of NE type, such as '*Born in 1929*' for PER type, but more contain mixed NE types.

Wikipedia categories are usually used in the method based on heuristic rules by defining category patterns [8, 24, 25]. Tardif et al. [18] directly tokenized the names of all categories applied to each article, and added each token to the BOW representation of the article. Consider that there are a large number of categories (approximately 140,000) in Chinese Wikipedia, but volunteers use only a few. Fig. 5 shows the percentage of categories that contain fewer than the specified number of articles. Approximately 21% of categories are empty and nearly 80% of categories contain fewer than 12 articles.

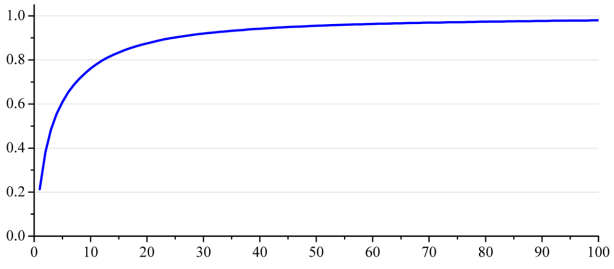


Fig. 5. Percentage of categories that contain less than the specified number of articles. The x-axis is the number of articles in each category, and the y-axis is the percentage of categories in the total.

Based on the analysis above, several constraints are designed for category selection. For a set of categories $O = \{o_1, o_2, ..., o_n\}$ that occurred in the training dataset, we consider each category o_i as an m-dimensional vector $(n_{i1}, n_{i2}, ..., n_{im})$, where n_{ij} is the number of articles that are contained by category o_i and are tagged with the j-th (j = 1, ..., m; m = 18) fine-grained NE type. For each category o_i, three constraints are set to select some representative categories.

Universality Constraint: A large number of categories are rarely used in Wikipedia. To filter less-used categories, we reserve the categories that contain more than two articles ($\sum_j n_{ij} > 2$) in the training dataset.

Centrality Constraint: The articles in the categories need to be annotated with a few NE types, such that the number that satisfies the condition $n_{ij} > 0$ (j = 1, 2, ..., m) is smaller than a third of the total number.

Superiority Constraint: The NE type can be considered as superior, if it has a more prominent number than others in the category. We use the variance of vector $(n_{i1}, n_{i2}, ..., n_{im})$ to measure superiority and leave the top 200 categories ordered by variance.

After category selection, a number of representative categories are selected. We build another feature vector based on the reserved categories, and add label 'Category' for all features, similar to the process employed for infobox subjects.

D. Article Title Feature

The article titles in Chinese Wikipedia provide internal evidence of NE type. For example, the article title that ends with '省' (*province*), is likely to be the name of GPE. Previous researches had paid little attention to article title, probably because these previous works had mainly handled English Wikipedia, which has an unobvious indicator of NE types. Tardif et al. [18] only tokenized the article title and added the tokens to the BOW representation of the article. For Japanese Wikipedia, Higashinaka et al. [19] refined 16 features from article titles including unigram or bigram, last common noun, and so on.

Many NER methods utilize the surface form of article titles and their context to extract and classify NEs. We design several NER features based on the structural characteristics of Chinese NEs.

1) NER Feature

Related researches on NER mainly focus on three universal NE types (PER, ORG, and LOC), for which some special features are usually designed. We give several special features that are based on Chinese NER.

Family Names: Some common Chinese family names,

such as '王' (*Wang*) and '赵' (*Zhao*), are typically used at the beginning of a Chinese name.

Number of Characters: Chinese names usually have two to four Chinese characters, including family name and given name.

Special Separator: The separator '•' is commonly used in Chinese names translated from foreign names, such as '马丁•路德•金' (*Martin Luther King*).

Common GPEs: Article title of ORG type is most likely to contain common GPEs, such as '*China Central Television*'; and the next is LOC type.

Last Chinese Character: The last Chinese character, such as '省' (*province*), '河' (*river*), is indicative of NE type.

2) Parenthesized Text

In addition to NER feature, another noticeable feature is the head nouns of parenthesized text in the article title. Approximately 10% of articles in Chinese Wikipedia are tagged with parenthesized text, which is mainly used to provide an explanation for a Wikipedia article. For example, '*novel*' and '*film*' in article titles '*David Copperfield (novel)*' and '*Topaz (1969 film)*', are both excellent indicators for the "Work of Art" NE type.

For all head nouns (terms) in parenthesized texts of article titles, we compute the frequency of each term and rank all terms by this frequency. Every 100 terms are grouped together from ranked terms, and the frequency sum of each group is computed. The bar-chart in Fig. 6 gives the percentage of the frequency sum of each group in the total frequency of all terms, whereas the line-chart in Fig. 6 shows the percentage of the cumulative frequency of top groups. Through statistical analysis for parenthesized text, we found that a significant coverage scale can be obtained by using only a small number of frequent terms. A large percentage of GPE terms can be found in parenthesized text (approximately 29% in total frequency for the top 100 terms), such as '*North District (Hong Kong)*' and '*George III (English)*'. To avoid the problem of feature sparsity, we filtered the low-frequency terms and transformed GPE and time expressions into generalization forms, such as '*Hong Kong*' to '*GPE*' and '*2013*' to '*Time*', in article titles '*North District (Hong Kong)*' and '*Tomb Raider (2013)*'. We then extracted head nouns, using simple part-of-speech modes.

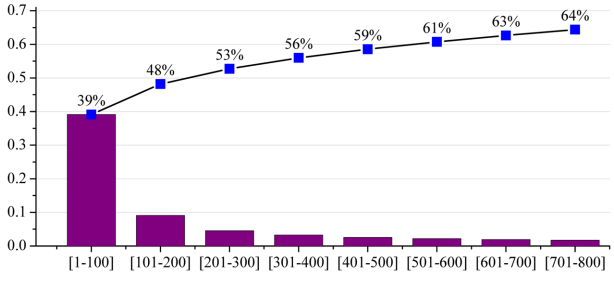


Fig. 6. Distribution information of the head nouns in parenthesized texts of article titles.

In this paper, we constructed several common gazetteers that could easily be collected on the Internet. These gazetteers include Chinese family names gazetteer, common GPE gazetteer, and indicative last Chinese character gazetteer. Then we built a new feature vector based on article title. This vector contains four special features (including family names, number of characters, special separator, and common GPEs), a BOW representation of the feature 'last Chinese character', and a BOW representation of the feature 'head nouns in parenthesized text'. The weights of the related feature are set to 1, if the current article title satisfies corresponding criteria; otherwise, the weights are set to 0.

V. EXPERIMENTS

A. Dataset

The Chinese version of Wikipedia (http://download.wikipedia.com/zhwiki/) in March 2013 is used in our experiments. This version contains 735,000 Wikipedia articles, and each article corresponds to a single Web page.

To create the dataset of Wikipedia articles with fine-grained NE types, two different sampling methods can be used: random sampling and popular sampling. Random sampling produces many Wikipedia articles that rarely occur in text, and only have small potential in NE applications. Moreover, these articles are more challenging, because they are rarely edited and some types of information are absent in Wikipedia. Popular sampling tends to choose Wikipedia articles targeted by the most frequent outgoing links. However, these articles usually refer to some common locations or concepts, such as '*China*' and '*Plant*' in Chinese Wikipedia. For some NE types (such as Celestial, Period, etc.), it is difficult to obtain sufficient instants from the articles targeted by frequent outgoing links.

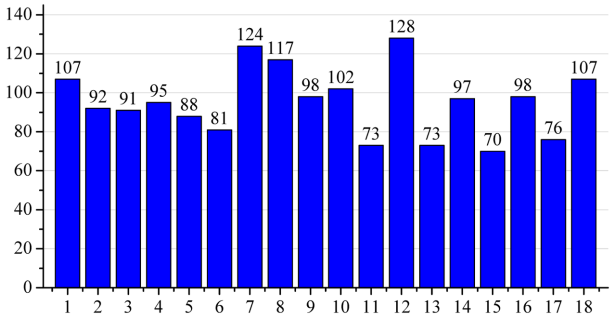


Fig. 7. The distribution of named entity (NE) types in the dataset. The x-axis represents the ID of fine-grained NE types, and the y-axis represents the number of each NE type that occurs in the dataset.

We adopted a compromise method that balances randomness and popularity in our experiments. Beginning from several seed articles, we randomly selected some articles occurring in the page of seed articles and then continued this process with selected articles. By using this method, we developed a dataset of 1,717 Chinese Wikipedia articles to evaluate the performance of NE classification. Fig. 7 shows the distribution of NE types. Two independent annotators were involved in the annotation process. For each article, the URL that points to the Web page of the Wikipedia article is generated. The annotators can more easily make their judgment about NE types of articles by using the richer information shown on the Web page.

B. Experimental Settings

Previous researches have mostly focused on English Wikipedia, and an important heuristic rule such that article titles are capitalized if they are proper nouns can be used to determine whether the articles are NEs [9]. However, this heuristic rule cannot be applied to Chinese Wikipedia. To distinguish NE type of non-entity (including common concepts and NE types not in our list) from other NE types, it must be added to the classifier.

The experiments adopted the strategy of two-layer classification. First, the types of LIST and DAB were easily identified, using some heuristic rules. The article titles of "List of …" pages usually end with the word '*列表*' (*list of …*). The article titles of the disambiguation pages usually contain '*消岐义*' (*disambiguation*) in parenthesized text, and the articles belong to the category '*Disambiguation*'. Then, a multi-classifier was trained based on the given features. All experiments were conducted with the SVM algorithms by using the toolkit libSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) with linear kernels, which presented excellent performance in [14, 16].

We evaluated the models by using 5-fold cross-validation and adopted the widely used Precision, Recall, and F1 to measure classification performance. The weighted average of each measure was adopted to evaluate the overall performance of all NE types. The weighted average of precision in t-th cross-validation (t = 1, 2, ..., 5), as well as those of Recall and F1, was computed using the following method.

$$precision_{avg}(t) = \frac{\sum_i (|c_i| \times precision(c_i, t))}{\sum_i |c_i|}$$

where $|c_i|$ is the number of the articles in NE type c_i (i = 1, 2, ..., m), and $precision(c_i, t)$ is the precision of NE type c_i in t-th cross-validation.

The global measurements of each NE type evaluated by 5-fold cross-validation were computed by:

$$precision(c_i) = \frac{1}{T}\sum_{i=1}^{T} precision(c_i, t)$$

$$precision_{avg} = \frac{1}{T}\sum_{i=1}^{T} precision_{avg}(t)$$

where T is the total number of cross-validation (T = 5). We adopted a similar method to compute $recall(c_i)$, $F1(c_i)$, $recall_{avg}$, and $F1_{avg}$.

C. Results and Discussion

In the experiments, the first-layer classification easily achieved outstanding performance by using the defined heuristic rules. Thus we emphasized the classification of other fine-grained NE types. As a common feature set, article content feature can effectively overcome the problem of zero-vector caused by other missing feature sets. We thus took the model applying only the article content feature as the baseline.

We performed several experiments as follows: 1) we observed the influence of feature dimension using the feature selection method of IG; 2) we verified the effectiveness of each feature set; and 3) we tested the performance of mixed feature sets.

Influence of Feature Dimension: High feature dimension consumes more computational resources with only a negligible improvement in classification performance. Fig. 8 shows the F1-measure for the article content feature set (baseline) of different feature dimensions, and we can see that sufficient representative features are covered, when the feature dimension reaches a particular value. In this work, we set the value of feature dimension to 3,000 and applied this parameter to the following experiments.

Effectiveness of Each Feature Set: To verify the effectiveness of each feature set, we trained classifiers by using the combination of each feature set and article content feature set (baseline), respectively, which are expressed as '+S' (structured feature set), '+C' (category feature set), and '+A' (article title feature set). Table 2 shows the experimental results. By analyzing the results, we can find some valuable conclusions as follows:

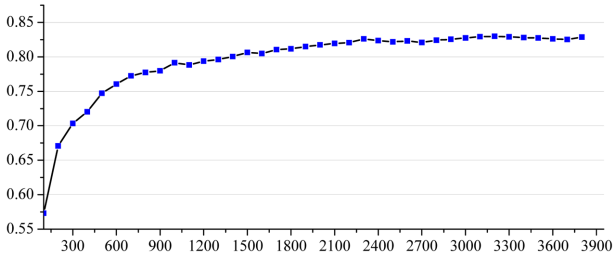


Fig. 8. F1-measure for article content feature set (baseline) of different feature dimensions (the step length is set to 100). The x-axis represents feature dimension, and the y-axis represents F1.

Table 2. Precision (P), recall (R) and F1-measure for each feature set

NE type		Baseline			+S			+C			+A		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
PER	People	75.6	62.0	67.7	78.2	80.3	78.9	87.2	80.3	83.0	67.1	85.1	74.7
	Non-People	55.4	70.1	61.8	80.4	71.8	75.7	90.3	53.4	66.7	77.0	55.6	63.9
ORG	Government	82.5	85.7	83.9	88.7	83.6	85.9	89.4	81.3	85.0	85.1	83.5	83.9
	Company	81.1	86.4	82.1	91.5	84.2	86.7	87.9	80.0	82.3	89.4	85.3	86.7
	Education	96.6	98.0	97.2	100	96.5	98.2	100	95.4	97.7	100	90.8	95.1
	Media	82.7	86.6	84.0	89.8	90.1	89.7	87.2	88.8	87.6	95.0	83.9	88.7
	Other	66.3	70.5	67.9	81.6	71.9	76.3	72.7	74.3	73.2	65.4	76.8	70.2
LOC	GPE	84.6	80.2	82.3	89.3	88.0	88.5	83.5	84.6	84.0	85.1	92.3	88.5
	Celestial	99.0	97.1	98.0	98.0	96.9	97.4	98.0	96.9	97.4	99.0	97.9	98.5
	Water Body	85.3	86.5	85.5	89.7	89.2	89.3	89.1	88.3	88.3	97.7	83.2	89.6
	Other	69.7	80.9	74.0	86.4	76.5	80.9	83.1	75.3	78.8	89.2	79.3	84.0
MISC	Facility	82.1	78.9	80.3	85.2	91.4	88.0	80.6	89.8	84.9	95.8	89.2	92.3
	Weapon	89.1	88.3	88.1	92.4	87.7	89.8	93.8	86.3	89.5	93.8	86.4	89.4
	Work of Art	73.1	69.1	70.6	77.2	80.4	78.3	66.8	83.4	73.7	73.6	75.1	74.0
	Period	88.6	93.3	90.5	93.4	88.6	90.5	95.9	88.6	91.8	97.3	87.1	91.5
	Language	99.0	99.0	99.0	99.0	99.0	99.0	99.0	98.0	98.5	100	98.0	99.0
	Game	94.8	87.0	90.5	91.5	97.4	94.4	92.3	93.4	92.8	90.2	93.4	91.5
NON	NON	75.7	80.6	78.0	68.4	87.0	76.3	66.0	86.8	74.9	61.7	76.8	68.4
Weighted average		83.3	82.3	82.3	**87.3**	**86.5**	**86.5**	86.0	84.6	84.5	85.9	84.4	84.6

NE: named entity, PER: person, ORG: organization, LOC: location, MISC: other, NON: non-entity, GPE: geo-political entity.

1) The addition of the feature set can improve performance at different levels. The combined feature set '+S' achieved better improvement of the F1-measure than the '+C' and '+A' feature sets.

2) For two fine-grained NE types in PER type (People and Non-People), the classifier performed worse while only using the article content feature set. The possible reason is that various descriptions exist in Wikipedia to introduce different people. The classification performance achieved great improvement while each feature set, especially structured feature set, was combined with article content feature set (+S, +11.2% for People and +13.9% for Non-People; +C, +15.3% for People and +4.9% for Non-People).

3) A certain feature set is a great indicator of some NE types. For example, the combined feature set '+A' could achieve great improvement for MISC:Facility (+12.0% F1-measure), LOC:Other (+10.0% F1-measure), and LOC:GPE (+6.2% F1-measure). After analyzing the evaluation data of these NE types above, we find that a lot of article titles have the same last Chinese characters, such as '*路*' (involving *road*, *railway*, etc.) and '*园*' (involving *park*, *garden*, etc.) for MISC:Facility, and '*岛*' (*island*) and '*山*' (*mountain*) for LOC:Other.

4) For part of NE types, such as MISC:Language and LOC:Celestial, outstanding performance could be achieved by the baseline system, because there are greater differences between the terms in article contents of these NE types. Furthermore, the NE type ORG:Other and LOC:Other achieved poor performance because they are composed of multiple fine-grained NE types.

5) For most fine-grained NE types, the systems trained by combined feature sets outperformed the baseline system by F1-measure. However, F1-measure of NON:NON type decreased because of poor precision (more articles belonging to each NE type are identified as non-entity type). The performance of combined feature set '+A' seriously decreased (-9.6% F1-measure), which resulted in minor improvement of overall performance by using this feature set '+A'.

Performance of Mixed Feature Sets: The mixed feature sets were used to identify the best combination of feature sets. We combined structured feature set (S), category feature set (C), and article title feature set (A) based on the baseline, and Table 3 reports the results of these mixed feature sets.

For different languages, it is difficult to perfectly implement the previous works because language-dependent

Table 3. F1-measure for mixed feature set

NE type		Baseline	+S+C	+S+A	+C+A	+S+C+A	Tardif et al. [18]	Tkatchenko et al. [17]
PER	People	67.7	89.9	81.6	75.0	81.6	90.8	87.7
	Non-People	61.8	76.2	75.0	72.0	72.9	80.6	79.2
ORG	Government	83.9	86.1	77.8	87.1	88.1	85.4	85.3
	Company	82.1	87.6	91.4	89.4	89.7	88.6	87.3
	Education	97.2	98.6	97.1	92.2	92.6	94.0	93.9
	Media	84.0	90.0	89.7	89.9	92.1	89.5	88.5
	Other	67.9	72.8	66.7	71.7	74.9	72.3	75.0
LOC	GPE	82.3	90.4	93.3	91.9	92.3	72.5	88.6
	Celestial	98.0	97.5	99.8	98.7	98.2	88.4	90.5
	Water Body	85.5	89.4	91.9	97.4	88.4	90.2	88.9
	Other	74.0	87.2	85.7	90.9	87.4	84.3	87.3
MISC	Facility	80.3	85.6	92.0	91.8	92.2	81.6	80.6
	Weapon	88.1	85.3	93.8	80.4	93.8	83.5	84.3
	Work of Art	70.6	82.7	78.9	82.3	79.2	74.3	71.4
	Period	90.5	92.3	89.7	92.1	96.3	92.1	91.1
	Language	99.0	98.7	95.0	98.6	95.4	93.5	93.1
	Game	90.5	90.6	96.6	93.5	93.5	86.2	87.7
NON	NON	78.0	76.3	66.7	72.7	65.0	63.6	61.7
Weighted average		82.3	**87.2**	86.4	86.8	86.9	83.3	84.0

NE: named entity, PER: person, ORG: organization, LOC: location, MISC: other, NON: non-entity, GPE: geo-political entity.

features and heuristic rules are usually adopted to achieve better classification performance, such as in Japanese [19] and Arabic [20]. In order to evaluate the resulting training set in classification, we re-implemented the state-of-the-art method presented by Tkatchenko et al. [17], and their baseline method that is similar to Tardif's classifier [18]. Their baseline method used the text of the first paragraph as a basic feature space, and a range of additional ones, namely, Title, Infobox, Sidebar, and Taxobox tokens, stemmed and tokenized categories, and template names. Category, Template, and Infobox features are extended with different prefixes (e.g., 'cat:') to distinguish them from the same tokens found in article text. Tkatchenko et al. [17] extended the features presented by Tardif et al. [18]. They added the 'List of' feature, which is constructed by tokenizing and stemming the titles of all 'List of …' articles containing the current article to the BOW representation. We neglected the boolean feature, which is used to increase the separability between articles about NEs and non-entities, because some English heuristic rules (such as Wikipedia naming conventions) were utilized to compute this feature, and pre-defined category patterns and template patterns could not be obtained.

In Table 3, we can see that the best performance was achieved by the mixed feature set '+S+C' (approximately 4.9% F1-measure higher than the baseline system). In all fine-grained NE types, significant improvements were achieved by People and Non-People in PER type, such as +22.2% F1-measure for People and +14.4% F1-measure for Non-People, by using mixed feature set '+S+C'. Moreover, the methods of Tardif et al. [18] and Tkatchenko et al. [17] outperformed the baseline system by 1%–2% F1-measure, but the overall performance of their methods in Chinese Wikipedia was far worse than the results of corresponding experiments in English Wikipedia.

From the results of Tables 2 and 3, we also find that the article title feature set can improve the classification performance, but the improvement is less significant than that of other feature sets. The reason is that this feature set mainly focuses on coarse-grained NE types and lacks discriminatory capability between fine-grained NE types. However, these coarse-grained NE types, especially three universal NE types (PER, ORG, and LOC), are important in many NLP applications. We also constructed training and test datasets with coarse-grained NE types, and the results in Table 4 show that a significant improvement could be achieved while adding the article title feature set (+S+C+A, +6.6% F1-measure; +C+A, +5.8% F1-measure; +S+A, +5.2% F1-measure).

Table 4. F1-measure for coarse-grained NE types

NE type	Baseline	+S	+C	+A	+S+C	+S+A	+C+A	+S+C+A	Tardif et al. [18]	Tkatchenko et al. [17]
PER	69.4	88.2	85.9	81.6	92.5	84.2	86.5	88.4	85.1	85.3
ORG	91.3	93.2	93.3	96.5	92.7	95.4	95.1	94.9	92.9	93.4
LOC	94.3	94.2	92.1	94.5	94.2	96.4	95.9	96.2	88.3	92.2
MISC	80.9	82.9	80.3	85.0	83.8	84.1	85.5	86.8	85.4	85.7
NON	61.1	82.0	80.3	79.6	78.8	79.0	80.9	82.0	48.9	51.3
Weighted average	84.3	88.3	86.7	89.4	88.8	89.5	90.1	**90.9**	85.9	87.1

NE: named entity, PER: person, ORG: organization, LOC: location, MISC: other, NON: non-entity.

VI. CONCLUSIONS

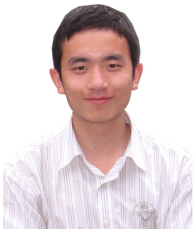
This paper presented a study on the NE classification of the articles in Chinese Wikipedia with fine-grained NE types. We explored four feature sets, namely, article content feature, structured feature, category feature, and article title feature. For each feature set, different strategies were designed to realize feature fusion. Experimental results showed that the combination of feature sets could effectively improve classification performance.

In future works, we will conduct experiments with hierarchical NE types using the method of hierarchical classification. Furthermore, the strategy of model fusion will combine the virtues of different models, such as the classifier and graph model. For application, the proposed method in this paper will be used to automatically generate Chinese NER corpora with the domain characteristics.

REFERENCES

1. J. Giles, "Internet encyclopaedias go head to head," *Nature*, vol. 438, no. 7070, pp. 900-901, 2005.
2. J. Kazama and K. Torisawa, "Exploiting Wikipedia as external knowledge for named entity recognition," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic, 2007, pp. 698-707.
3. H. Ji, R. Grishman, and H. T. Dang, "Overview of the TAC2011 knowledge base population track," in *Proceedings of the 4th Text Analysis Conference*, Gaithersburg, MD, 2011.
4. J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia," *Artificial Intelligence*, vol. 194, pp. 28-61, 2013.
5. J. Kazama and K. Torisawa, "Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations," in *Proceedings of Association for Computational Linguistics: Human Language Technologies*, Columbus, OH, 2008, pp. 407-415.
6. R. Bunescu and M. Pasca, "Using encyclopedic knowledge for named entity disambiguation," in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 2006, pp. 9-16.
7. D. M. Nemeskey and E. Simon, "Automatically generated NE tagged corpora for English and Hungaria," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Korea, 2012, pp. 38-46.
8. J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from Wikipedia," *Artificial Intelligence*, vol. 194, pp. 151-175, 2013.
9. J. Nothman, J. R. Curran and T. Murphy, "Transforming Wikipedia into named entity training data," in *Proceedings of the Australasian Language Technology Workshop*, Tasmania, Australia, 2008, pp. 124-132.
10. W. Chen, Y. Zhang, and H. Isahara, "Chinese named entity recognition with conditional random fields," in *Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing*, Sydney, Australia, 2006, pp. 118-121.
11. Y. Feng, L. Sun, D. Zhang, and W. Li, "Study on the Chinese named entity recognition using small scale character tail hints," *Acta Electronica Sinica*, vol. 36, no. 9, pp. 1833-1838, 2008.
12. A. Toral and R. Munoz, "A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia," in *Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 2006, pp. 56-61.
13. A. E. Richman and P. Schone, "Mining Wiki resources for multilingual named entity recognition," in *Proceedings of the Association for Computational Linguistics: Human Language Technologies*, Columbus, OH, 2008, pp. 1-9.
14. W. Dakka and S. Cucerzan, "Augmenting Wikipedia with named entity tags," in *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, Hyderabad, India, 2008, pp. 545-552.
15. Y. Watanabe, M. Asahara, and Y. Matsumoto, "A graph-based approach to named entity categorization in Wikipedia using conditional random fields," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic, 2007, pp. 649-657.
16. I. Saleh, K. Darwish, and A. Fahmy, "Classifying Wikipedia

articles into NE's using SVM's with threshold adjustment," in *Proceedings of the 2010 Named Entities Workshop*, Uppsala, Sweden, 2010, pp. 85-92.
17. M. Tkatchenko, A. Ulanov, and A. Simanovsky, "Classifying Wikipedia entities into fine-grained classes," in *Proceedings of IEEE 27th International Conference on Data Engineering Workshops*, Hannover, Germany, 2011, pp. 212-217.
18. S. Tardif, J. R. Curran, and T. Murphy. "Improved text categorisation for Wikipedia named entities," in *Proceedings of the Australasian Language Technology Association Workshop*, Sydney, Australia, 2009, pp. 104-108.
19. R. Higashinaka, K. Sadamitsu, K. Saito, T. Makino, and Y. Matsuo, "Creating an extended named entity dictionary from Wikipedia," in *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India, 2012, pp. 1163-1178.
20. F. Alotaibi and M. G. Lee, "Mapping arabic Wikipedia into the named entities taxonomy," in *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India, 2012, pp. 43-52.
21. D. Nadeau, P. D. Turney, and S. Matwin, "Unsupervised named-entity recognition: generating gazetteers and resolving ambiguity," in *Advances in Artificial Intelligence, Lecture Notes in Computer Science volume 4013*, Heidelberg: Springer, 2006, pp 266-277.
22. A. Sumida, N. Yoshinaga, and K. Torisawa, "Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in Wikipedia," in *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008, pp. 2462-2469.
23. J. H. Oh, K. Uchimoto, and K. Torisawa, "Bilingual co-training for monolingual hyponymy-relation acquisition," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Singapore, 2009, pp. 432-440.
24. L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proceedings of the 13th Conference on Computational Natural Language Learning*, Boulder, CO, 2009, pp. 147-155.
25. C. Bohn and K. Norvag, "Extracting named entities and synonyms from Wikipedia," in *Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications*, Perth, Australia, 2010, pp. 1300-1307.



Jie Zhou

Jie Zhou received his B.S. degree in information engineering, and M.S. degree in signal analysis and processing from Zhengzhou Information Science and Technology Institute, in 2007 and 2010, respectively. He has been working on theoretical research and experimental application of text processing technologies. The main research in his Ph.D. program is around sentiment orientation analysis, multimedia annotation, and named entity disambiguation.

Bicheng Li

Bicheng Li teaches at Zhengzhou Information Science and Technology Institute as a professor and Ph.D. supervisor. His teaching subjects include pattern recognition and artificial intelligence, and wavelet transform. He has published many professional books, such as in information fusion and application, and pattern recognition principle and application. His research fields include text analysis and understanding, speech/image/video processing and recognition, and information fusion.

Yongwang Tang

Yongwang Tang received his M.S. degree in intelligent information processing, and M.S. degree in control theory and control engineering from Kunming University of Science and Technology, in 2005 and 2008, respectively. His research fields include semantic Web technologies, information fusion, and database management.