Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Automatically building large-scale named entity recognition corpora from Chinese Wikipedia*

Jie ZHOU†, Bi-cheng LI, Gang CHEN

(*Department of Signal Analysis and Information Processing, Zhengzhou Information Science and Technology Institute, Zhengzhou 450002, China*)

†E-mail: zhoujie.nlp@gmail.com

Received Mar. 7, 2015; Revision accepted Aug. 9, 2015; Crosschecked Oct. 15, 2015

Abstract: Named entity recognition (NER) is a core component in many natural language processing applications. Most NER systems rely on supervised machine learning methods, which depend on time-consuming and expensive annotations in different languages and domains. This paper presents a method for automatically building silver-standard NER corpora from Chinese Wikipedia. We refine novel and language-dependent features by exploiting the text and structure of Chinese Wikipedia. To reduce tagging errors caused by entity classification, we design four types of heuristic rules based on the characteristics of Chinese Wikipedia and train a supervised NE classifier, and a combined method is used to improve the precision and coverage. Then, we realize type identification of implicit mention by using boundary information of outgoing links. By selecting the sentences related with the domains of test data, we can train better NER models. In the experiments, large-scale NER corpora containing 2.3 million sentences are built from Chinese Wikipedia. The results show the effectiveness of automatically annotated corpora, and the trained NER models achieve the best performance when combining our silver-standard corpora with gold-standard corpora.

Key words: NER corpora, Chinese Wikipedia, Entity classification, Domain adaptation, Corpus selection
doi:10.1631/FITEE.1500067 **Document code:** A **CLC number:** TP391

1 Introduction

Named entity (NE) is one of the basic language units, and plays an important role in natural language processing (NLP) tasks, such as text classification, question answering, and event extraction. Named entity recognition (NER) is aimed to identify the boundaries of named entities (NEs) in unstructured text and determine their NE types. These NE types have different definitions in different applications, such as DNA and RNA in molecular biology, product name and attribute in industrial manufacture. Generally, the most universal NE types refer to person (PER), organization (ORG), and location (LOC).

Supervised machine learning methods have become the currently dominant methods of NER. Unfortunately, an essential work for these methods is to manually annotate a large amount of text with linguistic information, which is a time-consuming task and requires significant skill. Current manually annotated data for machine learning is limited to several public datasets, almost exclusively newswire corpora on universal NE types. For different languages and predefined NE types, the corresponding annotated corpora are required to train new NER models. Furthermore, the models trained on a particular domain tend to perform worse on unseen domains. This data dependence has impeded the adaptation or porting of existing NER systems to new domains (Nothman *et al.*, 2013).

Recently, substantial research has been devoted to applying available online resources to NLP tasks, such as Linked Open Data (Ni *et al.*, 2010), Wikipedia

* Project supported by the National Natural Science Foundation of China (No. 14BXW028)
ORCID: Jie ZHOU, http://orcid.org/0000-0001-5615-9334
© Zhejiang University and Springer-Verlag Berlin Heidelberg 2015

(Medelyan *et al.*, 2009), and Wiktionary (Zesch *et al.*, 2008). The combination of these resources and NE technologies has been considered in some shared evaluations, such as the entity linking (EL) subtask of knowledge base population (Ji *et al.*, 2011). Compared with other resources, Wikipedia is the largest online collaborative multilingual encyclopedia and contains a larger number of NEs with more normalized structures. Wikipedia has been applied in many NLP applications, such as entity disambiguation (Ji *et al.*, 2011), entity gazetteers (Toral *et al.*, 2012), and semantic relatedness (Gabrilovich and Markovitch, 2007).

By leveraging existing human effort in Wikipedia's markup, Wikipedia is considered an appropriate resource to derive automatically annotated NER corpora for training machine learning models (Mika *et al.*, 2008; Nothman *et al.*, 2008; Richman and Schone, 2008). In this paper, millions of sentences in Chinese Wikipedia are transformed into NER corpora by identifying NE types of outgoing links and finding implicit entity mentions, e.g., the sentences shown in Fig. 1. We have built large-scale silver-standard corpora, which have lower quality than manually annotated gold-standard ones, but are suitable for training supervised NER systems for many more languages and domains (Nothman *et al.*, 2013). These corpora can be considered the initial annotations to build gold-standard corpora for predefined NE types, and this process will effectively reduce the workload of manual annotations. In some loose applications, they can directly replace gold-standard corpora to train NER models.

Chinese NER systems suffer from the same troubles on the NER corpora. Different from English NEs, Chinese NEs do not contain some important orthographic distinctions, e.g., the capitalization feature. Most previous research in this area focused on English language, and there is no specialized research on how to automatically build large-scale NER corpora from Chinese Wikipedia. In this paper, we refine novel and language-dependent features by exploiting the text and structure of Chinese Wikipedia, and automatically build enormous silver-standard NER corpora. Then we evaluate the performance of automatically annotated corpora. The main contributions are summarized as follows:

1. We combine four types of heuristic rules based on the characteristics of Chinese Wikipedia and the supervised NE classifier to identify NE types of Chinese Wikipedia articles accurately and comprehensively.

2. To avoid missed annotations caused by unlabeled links, we present a method to find implicit mentions in article content using boundary information of outgoing links, and identify NE type of ambiguous mentions based on the EL method.

3. We present a tagged corpus selection approach based on core article extending, which can automatically adapt the domains of the test data.

2 Related work

Named entity recognition, as one of the most important sub-tasks of information extraction, was first defined by the sixth Message Understanding Conference (MUC-6). Early research was mostly based on handcrafted rules. Most of recent methods use supervised machine learning, relying on an annotated training corpus, from which a learning algorithm infers patterns associated with NEs, based on morphological, syntactic, lexical, and contextual features.

2003年6月17日，[PER 贝克汉姆]与[LOC 西班牙][ORG 皇家马德里]签了一纸四年合约，总值 3500 万欧元（相当于 2500 万英镑或超过 4000 万美元）。	[Time 2003 年 6 月 17 日]，[PER 贝克汉姆]与[Country 西班牙][ORG 皇家马德里]签了一纸[Time 四年]合约，总值[Monetary 3500 万欧元]（相当于[Monetary 2500 万英镑]或超过[Monetary 4000 万美元]）。
On 17 June 2003, [PER Beckham] joined [LOC Spanish] champions [ORG Real Madrid] for 35 million euros (about 25 million pounds or more than 40 million dollars) on a four-year contract.	On [Time 17 June 2003], [PER Beckham] joined [Country Spanish] champions [ORG Real Madrid] for [Monetary 35 million euros] (about [Monetary 25 million pounds] or more than [Monetary 40 million dollars]) on a [Time four-year] contract.
(a)	(b)

Fig. 1 An example of transforming sentences in Chinese Wikipedia into annotated NER corpora (and English translation) for universal NE types (PER, ORG, and LOC) and fine-grained NE types (PER, ORG, Time, Country, Monetary, etc.): (a) annotation for universal NE types; (b) annotation for fine-grained NE types

The blue text corresponds to the outgoing link in Wikipedia. References to color refer to the online version of this figure

However, because of the data dependence of supervised machine learning methods, some work has focused on semi-supervised and unsupervised methods. Nadeau *et al.* (2006) proposed an NER system that combines NE extraction (inspired by the study of Etzioni *et al.* (2005)) with a simple form of NE disambiguation. In a comparison on the MUC corpus, their system outperforms a baseline supervised system, but it is still not competitive with more complex supervised systems. Liao and Veeramachaneni (2009) presented a semi-supervised learning algorithm for NER using conditional random fields (CRFs). They adopted an independent evidence to automatically extract high-accuracy and non-redundant data, and realized improvement for the classifier at each iteration. Liu *et al.* (2011) proposed to combine a *K*-nearest neighbors classifier with a linear CRFs model under a semi-supervised learning framework to solve the problem of insufficient information in a tweet and the unavailability of training data.

Current manually annotated data for machine learning is usually sampled from special domains. NER methods usually lose accuracy in the domain transfer due to the different data distribution between the source and the target domains (Ciaramita and Altun, 2005; Guo *et al.*, 2009). The major reason for performance degradation is that each entity type often has a lot of domain-specific term representations in the different domains (Guo *et al.*, 2009). It is impractical to build a manually annotated training dataset for every target domain. Jiang and Zhai (2006) automatically ranked features based on their generalizabilities across domains, and trained a classifier with strong emphasis on the most generalizable features. Then Jiang and Zhai (2007) presented a two-stage approach to domain adaptation, including recognizing the generalizable features and learning an appropriate weight with the use of a modified logistic regression framework. Guo *et al.* (2009) grouped words from the unlabeled corpus into a set of concepts according to the related context snippets, and projected the original term spaces of both domains to a concept space.

The need for manually annotated corpora limits the creation of high-performance NE recognizers for most languages and domains, so researchers try to use existing resources, such as Web (An *et al.*, 2003), cross-language resources (Ehrmann and Turchi, 2010; Fu *et al.*, 2011), and Wikipedia (Mika *et al.*, 2008; Nothman *et al.*, 2008; Richman and Schone, 2008), to automatically build NER corpora. Wikipedia collects a lot of existing human effort in the markup, which makes it an appropriate resource to build NER corpora.

Nothman *et al.* (2008) used a four-step treatment method, namely, entity classification of Wikipedia articles, sentence segmentation of article contents, entity labeling in text, and sentence selection. Then they conducted further research on building multilingual NER corpora (Nothman *et al.*, 2013). Richman and Schone (2008) used a method similar to the method used by Nothman *et al.* (2008) to build NER corpora in languages other than English. They classified Wikipedia articles in other languages by transferring knowledge from English Wikipedia via cross-language links. Mika *et al.* (2008) explored the use of infobox information, rather than outgoing links, to derive NE annotations. They extracted attribute–value pairs from infobox templates, and learned associations between NE types and infobox attributes by tagging English Wikipedia text with a CoNLL-trained NER system. This mapping is used to project NE types onto the labeled instances which are used as NER training data.

Most studies have used language-dependent characteristics, such as conventional capitalization in English (Nothman *et al.*, 2008), or knowledge bases, such as DBpedia (Nemeskey and Simon, 2012). For different languages, substantial language-dependent characteristics, which are not mentioned in related research, can be developed to improve performance effectively.

Compared with English NER systems, Chinese NER systems usually perform worse. Among the main NE resources available for Chinese are the manually annotated datasets from SIGHAN NER tasks. There are three corpora in the evaluation of SIGHAN 2006, including: corpus from the City University of Hong Kong, the Microsoft Research corpus, and the Linguistic Data Consortium corpus. Another important available resource is the Peking University corpus, which has been used in the evaluation of SIGHAN 2008. These corpora are built based on the annotation scheme of universal NE types, and there are few public available corpora with

fine-grained NE types at present. Chinese Wikipedia contains a large number of articles (approximately 735 000 as at March 2013), but currently related research combining NEs and Chinese Wikipedia mainly focuses on semantic relatedness computing (Liu and Chen, 2010) and entity relation extraction (Zhang *et al.*, 2012). Using this outstanding resource of Chinese Wikipedia, we take the first attempt to automatically build large-scale Chinese NER corpora.

3 Building Chinese NER corpora

Wikipedia contains a wealth of multi-faceted information, including articles, links between articles, categories, infoboxes, a hierarchy that organizes categories and articles into a large directed network, and cross-language links. Therefore, we exploit these various types of information for building NER corpora.

To obtain better NER corpora, we have to solve two problems: (1) building high-quality NER corpora and (2) training better NER models by using automatically annotated corpora. In this section, we present a method to build large-scale Chinese NER corpora automatically.

For each sentence in the article content of Chinese Wikipedia, if this sentence contains explicit mentions, which are expressed by outgoing link markup of square brackets [[entity|mention]], we tag them with NE types of referred entities determined by the entity classification of explicit mention. Then, implicit mentions in the sentence are identified and tagged with predefined NE types. Lastly, the sentence selection approach is used to adapt the domains of the test data. Fig. 2 shows an overview of the Chinese NER corpora generation process. The detailed description of each step is presented in the following.

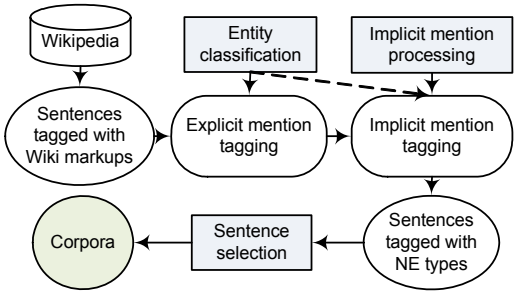

Fig. 2 An overview of the Chinese NER corpora generation process

In this paper, we adopt three universal NE types (PER, ORG, and LOC) that correspond to each type of the CoNLL annotation scheme (http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt), and classify other entity types (e.g., 'Auto' and 'Work of Art') and non-entities into OTHER type.

3.1 Entity classification of explicit mention

For Wikipedia articles, we can identify explicit mentions in article content using Wiki markup of outgoing links and build a map between explicit mention and its target article. Thus, the problem that determines NE types of these mentions can be transformed into entity classification of all articles. The high performance of entity classification is essential because classification errors can transfer into NER corpora. Currently, the work on entity classification has mainly focused on two types, heuristic rule based and supervised classifier based. The methods based on heuristic rules usually achieve high precision, but a large number of Wikipedia articles still cannot be covered by heuristic rules. To achieve high performance and large coverage, a method that combines heuristic rules of multi-faceted information with a supervised NE classifier is designed to determine the NE types of all articles in Chinese Wikipedia.

3.1.1 Heuristic rules of multi-faceted information

By constructing heuristic rules, we can precisely determine the NE types of Wikipedia articles. We adopt four types of heuristic rules, including Wikipedia category, infobox template, cross-language link, and article title. The vector $\boldsymbol{C}_x$ denotes the membership degree to four NE types (PER, ORG, LOC, and OTHER). The value of $\boldsymbol{C}_x$ is set to one if the article can match related rules. The comprehensive membership degree vector $\boldsymbol{C}_{\text{multi}}$ is computed by weighting each type of heuristic rule. Then the NE type with a unique maximum value is set as the final result:

$$\boldsymbol{C}_{\text{multi}} = \alpha_1 \boldsymbol{C}_{\text{category}} + \alpha_2 \boldsymbol{C}_{\text{infobox}} + \alpha_3 \boldsymbol{C}_{\text{language}} + \alpha_4 \boldsymbol{C}_{\text{title}},$$

where the weights satisfy the constraint $\sum_{i=1}^{4} \alpha_i = 1$.

Wikipedia category: Wikipedia provides its own category structure that is described as a folksonomy. This category structure is not a simple tree-structured taxonomy but a graph in which multiple schemes

coexist. The categories usually express a relation that is common to all articles in the category. Some relations are great indicators of NE types but most relations are not. The key task is to refine the heuristic rules of categories related to NEs and determine their NE types.

Richman and Schone (2008) searched the category hierarchy in English Wikipedia until a threshold of reliability was passed or a preset search limit was reached. Based on this research, Nothman *et al.* (2013) created detailed case-sensitive keywords or phrases that matched English Wikipedia categories for NE classification. Ratinov and Roth (2009) manually aggregated several categories into a higher-level concept including universal NE types, and added Wikipedia articles tagged by these categories to the corresponding NE gazetteer. These studies listed their heuristic rules of the category name in English Wikipedia.

Given that the category structure is created collaboratively by many volunteers, child categories may inherit different aspects of the parent category that changes the relation with NEs. For example, as the indicator of PER entity, category 'person' has child categories 'occupation' and 'character biography', which do not express a direct relation with NEs. To avoid the above-mentioned problem, we tag only the articles that are directly affiliated in the categories matching our heuristic rules. The application of this constraint results in higher precision at the expense of recall, and recall can be compensated using other heuristic rules.

Because of morphological and orthographic distinctions between different languages, heuristic rules need to be established based on the characteristics of each language. For example, the experts in all kinds of fields can be generalized by the phrase '…学家' in the Chinese Wikipedia category, such as '哲学家' (philosopher), '天文学家' (astronomer), and '数学家' (mathematician). We create a list of category patterns using some predefined expressions shown in the Appendix (Table A1). The disambiguation articles (DAB) are identified individually as a special type because the mention of name ambiguity can denote several distinct entities. Four customizable expressions are used in our category patterns, where '[*x*]' denotes an optional character '*x*', '[*]' denotes any (zero or more) character, '[+]' denotes at least one character, and '<TIME>' denotes time expression such as '19th century', 'the 1990s', or '1990'.

Infobox template: Infobox is a special type of template that often contains a condensed set of important facts relating to the article. As highly structured information, infoboxes in Wikipedia are often used to construct machine-readable datasets, such as DBpedia (Auer *et al.*, 2007) and WikiNet (Nastase and Strube, 2013).

Usually, Wikipedia editors tend to use the same infobox template with similar existing articles. For example, the infobox template 'Infobox football biography' most likely appears in Wikipedia articles 'David Beckham' and 'Ronaldo'. Thus, we assume that the articles are likely to have the same NE type while using the same infobox template. Our results have also proven this assumption.

We extract 1294 infoboxes from the page of Wikipedia category 'infobox template' and annotate each infobox with NE type or ambiguous type. English and Chinese keywords in the CoNLL annotation scheme are used as query conditions to determine the initial type. Thereafter, we check the results manually to obtain the final annotations. In fact, we need only to annotate a few of the most frequent infoboxes that have great coverage. When 500 most frequent infoboxes are selected, the coverage reaches over 99% in the articles that contain infobox templates (approximately 47.7% of the total).

Cross-language links: Wikipedia is a multilingual resource that involves more than 200 languages. The same articles in different languages are linked by special Wiki markup.

Following the naming conventions of proper names in English, article titles or related aliases in incoming links are capitalized if they are proper names. This heuristic rule can be used to predict the non-entity type (OTHER type) except for some conventional capitalization (Bunescu and Paşca, 2006; Nothman *et al.*, 2013).

Unlike English, some languages (e.g., Chinese) do not have obvious superficial features to distinguish NE types from non-entity type. A common way of entity classification in non-English Wikipedia is to find associated English articles and then determine NE types of non-English articles by that of the associated English articles (Richman and Schone, 2008; Nemeskey and Simon, 2012; Darwish, 2013).

In Chinese Wikipedia, there are a large number of non-entity articles that involve many domains, such as '生物病毒分类表' (virus classification), '自然语言' (natural language), and '中国武术' (Chinese martial arts). It is difficult to distinguish non-entities from these universal NE types using a supervised NE classifier. To improve the performance of entity classification, the heuristic rules of cross-language links are adopted to identify non-entity articles in Chinese Wikipedia by capitalized conventions of associated English articles. The following criteria are adopted to identify non-entity articles:

1. Wikipedia articles which comprise time expressions in the titles of associated English articles, such as '2007 French Super Series';

2. Wikipedia articles which have a larger or equal number of lowercase first characters than capitalized ones in the words of article titles, except several special words (such as 'the', 'of', 'de', 'no', and 'von').

In this way, 50 000 articles (approximately 6.8% of the total) in Chinese Wikipedia are considered to be non-entity by heuristic rules of cross-language links.

Article title: The title of Wikipedia article is a succinct and exclusive phrase that corresponds to a specific web page. Previous research had paid little attention to the article title probably because these previous studies had handled mainly English Wikipedia, which has an unobvious indicator of NE types. For Japanese Wikipedia, Higashinaka *et al.* (2012) refined 16 features from article titles, including unigram or bigram, last common noun, etc.

In Chinese Wikipedia, some last common nouns of article titles can be a great indicator of NE types, such as '[+]火车站' (railway station) and '[+]大学' (university). Using the last common nouns, we can also easily identify special types of articles, such as the 'List of …' page, and tag them with OTHER type.

Another noticeable term that can distinguish NE type is the parenthesized expression. Approximately 10% of articles in Chinese Wikipedia are tagged with parenthesized expressions to resolve the ambiguity between articles sharing a name. We create a list of article title patterns based on last common nouns and parenthesized expressions shown in the Appendix (Table A2). Each NE type may be influenced differently by its own characteristics. The PER type uses only parenthesized expressions, the ORG and LOC types use only last common nouns, and the OTHER type uses both.

3.1.2 Supervised named entity classifier

The method based on the heuristic rule can achieve high precision but has very limited coverage. To overcome this disadvantage, we need a method which can be used to determine the NE types of all articles in Wikipedia, including those that cannot match any of heuristic rules.

The supervised NE classifier is a widely used and effective method in multiple languages, such as English (Dakka and Cucerzan, 2008), Arabic (Alotaibi and Lee, 2012), and Japanese (Higashinaka *et al.*, 2012). We have explored four types of feature sets, namely, article content feature, structured feature, category feature, and article title feature. Then, 2000 articles are selected randomly and annotated manually as the training data. Finally, we have trained a support vector machine (SVM) classifier to classify the articles that have two or more equal maximum values for the comprehensive membership degree vector $\boldsymbol{C}_{\text{multi}}$ (including zero-vector).

Article content feature: Article content refers to a detailed description of the Wikipedia article, which introduces the related knowledge in textual format. It is usually considered a group of common features because almost all Wikipedia articles contain such information.

Unlike in 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 keep only the set of the most representative part-of-speech (noun, verb, and adjective). Then the feature vector that contains the bag-of-words (BOW) representation of reserved terms is built, and the feature weight of each word is computed using the term frequency–inverse document frequency (TF-IDF) method.

Although the feature space is effectively limited by part-of-speech selection, the vector dimension remains large, and the data could be plagued by a substantial amount of noise. To reduce the feature dimension, the feature selection method of information gain (IG) is introduced. The feature dimension is set to 2000 in the experiment.

Structured feature: The articles in Wikipedia are edited using Wiki markup language, and some important information can be tagged (e.g., section title,

infobox template). We discard outgoing links because they are more inclined to explain the topics of related articles. Finally, three representative structured features are chosen, namely, section title, infobox, and co-occurrence relation in the lists and tables.

1. Section titles are usually the framework of an article. Words in the section titles can indicate NE types more clearly, such as 'early life' for PER type. To combine with the feature vector of article content, we increase the weight of words contained in the section titles, except some common ones such as 'references' and 'links'.

2. Infobox templates are a set of subject-attributes-values triples that often contains a condensed set of important facts relating to the article. For example, in Wikipedia article 'Michael Jordan', the subject 'Infobox NBA biography' of infobox and many pairs of attribute-value, such as attribute 'birth place' and value 'Brooklyn, New York', are used to describe this article. Words contained in the values of infobox are processed using the same method as for section titles. Then, we build another feature vector based on the infobox subject, and add label 'Infobox' for all features, such as 'Infobox: Infobox NBA biography'. The weights are computed simply by examining the presence of the subject.

3. Some co-occurrence articles, which have strong semantic correlation and same NE type, can be extracted from the tables and lists in Wikipedia pages, especially in 'List of …' pages. For example, in Wikipedia article 'List of mountains', some articles of semantic correlation are listed, such as 'Everest', 'K2', and 'Kangchenjunga'. For the current article, we extract common words from section titles and infobox values of its co-occurrence articles, which can be seen as important features of these articles. To highlight the commonness, we retain only the frequent words which occur in half of the co-occurrence articles at least, and increase the weight of words in the feature vector of article content.

Category feature: As introduced in heuristic rules, some Wikipedia categories are great indicators of NE type, such as '1929 births' and '20th-century American writers' in Wikipedia article 'Martin Luther King', but more are not. We want to select some representative categories from a large number of ones in Chinese Wikipedia (about 140 000 categories).

For a set of categories $\boldsymbol{O}=\{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_n\}$ occurring in all Wikipedia articles of the training dataset, we consider each category $\boldsymbol{o}_i$ an m-dimensional vector $(n_{i1}, n_{i2}, \ldots, n_{im})$, where n_{ij} is the number of articles that contain category $\boldsymbol{o}_i$ and are tagged with the jth NE type (corresponding to PER, LOC, ORG, and OTHER). For each category $\boldsymbol{o}_i$, three constraints are set to select some representative categories:

1. 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.

2. 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 less than a predefined threshold. In our experiments, we set the threshold 1 because the articles need only to be classified into four NE types.

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

Then, 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.

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 a location. We design several NER features based on the structural characteristics of Chinese NEs:

1. Family names: Some common Chinese family names, such as '王' (Wang) and '赵' (Zhao), are typically used at the beginning of a Chinese name.

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

3. Special separator: The separator '•' is commonly used in Chinese translation of foreign names, such as '马丁•路德•金' (Martin Luther King).

4. Common geographical/social/political entities (GPEs): The article title of ORG type is most likely to contain common GPEs, such as 'China Central Television', and the next is LOC type.

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

In addition to the NER feature, another noticeable feature is the head nouns in parenthesized text in the article title. For example, 'novel' and 'actor' in article titles 'David Copperfield (novel)' and 'Adam Williams (actor)' are excellent indicators of OTHER and PER types, respectively. To avoid the problem of feature sparsity, we transform GPE and time expressions to the generalization forms, such as 'Hong Kong' to 'GPE', '2013' to 'Time' in article titles 'North District (Hong Kong)' and 'Tomb Raider (2013)', respectively.

We construct several common gazetteers that can be collected easily on the Internet, including: Chinese family names gazetteer, common GPE gazetteer, and indicative last Chinese character gazetteer. Then we build a new feature vector based on the article title. This vector contains four special features corresponding to the first four NER ones, a BOW representation of the last Chinese character, and a BOW representation of head nouns in parenthesized text. The weights of the related feature are set to 1 if the current article title satisfies the corresponding criteria; otherwise, they are set to 0.

3.2 Type identification of implicit mention

To identify the implicit mentions in the article content, we need to solve two problems: (1) identification of the boundaries of implicit mentions; (2) identification of NE type of ambiguous mentions.

3.2.1 Finding implicit mention with NE type

In Wikipedia article content, only the first occurrence of a particular mention is typically tagged with a corresponding outgoing link, whereas subsequent mentions are usually untagged. Some outgoing links to Wikipedia articles are overlooked because of their insignificant correlation with the current article or volunteers' negligence.

To build silver-standard corpora with high recall, we need to find implicit mentions of three universal NE types (PER, ORG, and LOC). The first task is to construct a query list to determine whether the extracted string is an implicit mention. The following query lists are constructed based on two different sources:

1. Global query list: This list contains all article titles and redirect titles of articles in Chinese Wikipedia, and the NE type of each title is tagged by a corresponding Wikipedia article. We build a global set of triples <title, Wikipedia article, NE type>.

2. Local query list: This list contains the anchor text of outgoing links in the current Wikipedia article, which is expressed by Wiki markup [[entity|mention]] and usually seen as the alias of the target article. Moreover, the first and last names in Chinese translation of foreign names, such as '大卫' (David) and '贝克汉姆' (Beckham) in '大卫·贝克汉姆' (David Beckham), are added to this list as short names. We also build a local set of triples <title, Wikipedia article, NE type>. For each article, a different local query list is constructed.

For all titles in the lists above, the canonical expressions are obtained by special processing, such as removing the underline and parentheses expressions, converting Chinese traditional characters to simplified ones. For example, article title '喬治三世_(英國)' (George III of the United Kingdom) needs to be converted to canonical expression '乔治三世' (George III). Then all titles in each list are expressed as a trie for fast query.

When continuous Chinese characters are composed in one query string, Chinese words may be split at the border of the query string; this problem is known as 'segmentation ambiguity'. Therefore, we realize Chinese word segmentation using the NLPIR toolkit and take Chinese words as the minimum unit, instead of Chinese characters. Although an outstanding performance of Chinese word segmentation is achieved using the NLPIR toolkit, some mistakes arise at the border of the anchor text of the outgoing link. In this case, further segmentation is conducted along the border. For example, the one-character word '令' and anchor text of outgoing link '李靖' are combined incorrectly in the process of Chinese word segmentation shown in Fig. 3. Thus, further segmentation is required at the border of Chinese word '李靖'.

In the article content of Chinese Wikipedia, outgoing links tag the explicit mentions and give their target articles. Furthermore, outgoing links contain the semantic information of processing boundaries. By considering this available information, we realize

the implicit mention finding algorithm that is described as follows:

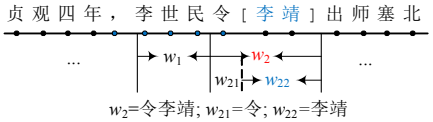

Fig. 3 Example of further segmentation to correct the word segmentation mistake

Algorithm 1 Implicit mention finding

Input: a set of sentences $\boldsymbol{S}=\{S_1, S_2, \ldots, S_{|\boldsymbol{S}|}\}$, where S_i is a sequence of Chinese words; for each sentence, there is a set of tuples $\boldsymbol{T}_{\text{explicit}}=\{(g_1, t_1), (g_2, t_2), \ldots, (g_p, t_p)\}$, where g_j is a segment tagged by the outgoing link and t_j is NE type identified by entity classification.

Output: a set of tuples $\boldsymbol{T}_{\text{implicit}}=\{(e_1, t_1), (e_2, t_2), \ldots, (e_q, t_q)\}$ for each sentence, where e_j is a segment of implicit mention and t_j is NE type of implicit mention.

For each $S_i \in \boldsymbol{S}$
  Split the sentence S_i into segments by the boundary of g_j (j=1, 2, …, p); // Fig. 4a
  If the segments do not belong to $\boldsymbol{T}_{\text{explicit}}$ **or** NE types are OTHER type **then**
    Add the segments to the set $\boldsymbol{Q}$; // Fig. 4b
  End If
  For each $g \in \boldsymbol{Q}$
    Set starting position sp←1;
    // length(g) expresses the length of segment g
    While sp≤length(g)
      Find the longest matching string e, starting with sp and corresponding to NE types using the forward maximum matching algorithm; // Fig. 4c
      If e exists **then**
        Add e and NE type t to $\boldsymbol{T}_{\text{implicit}}$;
        sp←sp+length(e)−1;
        sp←sp+1;
      End If
    End While
  End For
End For
Return $\boldsymbol{T}_{\text{implicit}}$ for each sentence;

3.2.2 Determining NE type of ambiguous mention

When determining the NE type of a mention, we can easily make judgment if the mention has only one referent article. However, a mention sometimes corresponds to more than one referent article. For example, the implicit mention ‘George’ may be any one of the following articles with the same name: ‘George Washington’, ‘George (Washington)’, ‘George (magazine)’, etc.

To process ambiguous mentions, we can map the mention to this referent article if the mention is exactly the same as the initial expression of the article title or redirect title (only converting Chinese traditional characters to simplified ones), because this name can be considered a popular one for this Wikipedia article. For other mentions, we need to build a mapping relation between the given mention m and NE types $\{c_{\text{per}}, c_{\text{org}}, c_{\text{loc}}, c_{\text{other}}\}$; this process is similar to the EL task. Hence, the obvious solution is to realize the mapping $m \rightarrow e_k$ ($e_k \in \{e_1, e_2, \ldots, e_m\}$) using EL technology and tag mention m with the NE type of article e_k. However, the accurate mapping between the given mention and the referent articles is not necessary in our task. For the articles with the same query name, we collect the articles of the same NE type as a new set and realize the mapping $m \rightarrow \boldsymbol{e}_k$ ($\boldsymbol{e}_k \in \{\boldsymbol{e}_{\text{per}}, \boldsymbol{e}_{\text{org}}, \boldsymbol{e}_{\text{loc}}, \boldsymbol{e}_{\text{other}}\}$), where $\boldsymbol{e}_k$ denotes a set of all articles tagged with one NE type.

Based on the method of context-article similarity in Bunescu and Paşca (2006), we determine the NE type of an ambiguous mention. The goal can be written as

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}_k} \text{score}(m, \boldsymbol{e}_k).$$

If the maximum score is less than the threshold, we tag the current mention with the OTHER type,

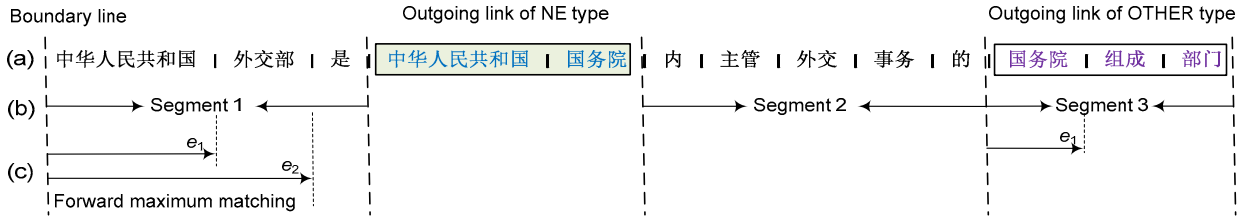

Fig. 4 Example of the implicit mention finding for a sentence

The sentence is composed of a sequence of Chinese words (separated by vertical lines). In (a), two outgoing links are tagged in the sentence (represented by boxes); in (b), the outgoing link of OTHER type is converted to text segment, and three segments (separated by the dotted line) are obtained in the sentence; in (c), the longest string is obtained using the forward maximum matching algorithm in segment 1

which has a similar processing to the NIL of EL. The sentences which contain the query name and its expanded form (short name, full name, alias, etc.) are selected as the context of the mention. The ranking function is based on the cosine similarity between the context of the mention and the text of all articles in the related set:

$$\text{score}(m, \boldsymbol{e}_k) = \cos(m.T, \boldsymbol{e}_k.T) = \frac{m.T}{\| m.T \|} \frac{e_k.T}{\| \boldsymbol{e}_k.T \|},$$

where $m.T$ and $\boldsymbol{e}_k.T$ denote the feature vectors computed using the classical method for weight computing, TF-IDF.

3.3 Tagged corpus selection approach

As one of the largest online repositories of encyclopedic knowledge, Wikipedia involves all kinds of domains that usually have different expressions and distributions of NEs, such as geography, biology, and entertainment. If all annotated data is used to train the NER model, the model performance may be influenced by a large amount of irrelevant data.

To select the corpus of more precisely annotated data, Nemeskey and Simon (2012) removed the sentences that contain an unknown NE type for the link target, and gave the users option to decide whether they wanted to use 'low quality' sentences. Nothman *et al.* (2013) selected portions of data in their corpus using criteria based on the confidence that all capitalized words are linked to the articles of known NE types. These methods tend to select high confident sentences as the NER corpus, but do not consider the application domain of an automatically annotated corpus.

The NER model is generally trained by gold-standard corpora, most of which come from newswire about national contemporary politics. However, this model may be applied to process documents from other domains in practice, thus resulting in poor performance. We design an approach based on core article extending to select data that is related to current domains. The process is as follows:

Step 1 (core article extraction): A sample set of test data is extracted (in our experiment, all test data is used), and implicit mentions are identified using the method discussed in Section 3.2. Mentions tagged with the PER or ORG type, which are the representative names of the domain, are then extracted, and the corresponding Wikipedia articles of the mentions above are added to the set of core articles $\boldsymbol{E}_{\text{core}}$ (if the mention has more than one correlative Wikipedia article, all articles are added to $\boldsymbol{E}_{\text{core}}$).

Step 2 (core article extending): The article set is extended using the target article of outgoing links in the text of each core article. Extended articles that do not contain the NE of the ORG type in the text are removed to overcome the problem of low ORG density in Chinese Wikipedia. The set of extended articles is denoted as $\boldsymbol{E}_{\text{ext},1}$.

Step 3 (article ranking): The articles in the set $\boldsymbol{E}_{\text{ext},1}$ are firstly ranked by the number of times that they occur in the core articles as the target article of the outgoing link, followed by the length proportion of NE mention text and article text in each article.

Step 4 (corpus building): Sentences are extracted from ranked Wikipedia articles in order, and the set $\boldsymbol{E}_{\text{core}}$ is considered superior to $\boldsymbol{E}_{\text{ext},1}$. An additional criterion that requires the sentence contain at least one tagged NE is applied. The process above is continued until a specified number of sentences or tokens are added to the corpus.

4 Experiments

The Chinese version of Wikipedia in March 2013 is used in our experiments. This version contains 735 000 Wikipedia articles, and each article corresponds to a single web page.

4.1 Evaluation of entity classification

The entity classification performance has a significant effect on the quality of automatically annotated corpora because classification errors can transfer into these corpora. We first evaluate the entity classification approach in the experiments.

To classify Chinese Wikipedia articles into NE types, we create four types of heuristic rules that use multi-faceted information in Chinese Wikipedia. The proportions of matched articles on the four types of heuristic rules are shown in Fig. 5. About 66% articles in total can match at least one rule and the heuristic rules of the infobox and category have greater coverage of articles than those of the article title and cross-language link. Furthermore, there are many Wikipedia articles that can match multiple types of heuristic rules. For example, in Wikipedia articles,

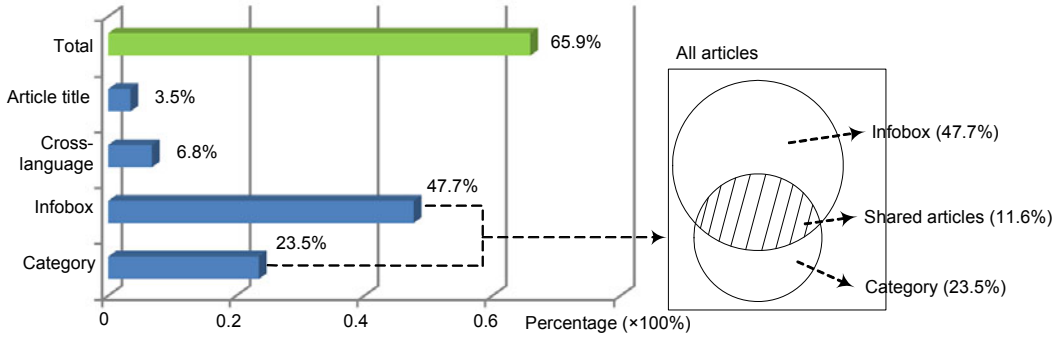

Fig. 5 Proportions of matched articles by the heuristic rules of multi-faceted information

the percentages of articles matched by infobox rules and category rules are 47.7% and 23.5%, respectively. We find that about half of articles in category rules (11.6% in total) are also contained in infobox rules.

Dataset: In our experiments, a dataset of 2000 Chinese Wikipedia articles is developed as the training data (500 articles each for PER, ORG, LOC, and OTHER types). Because the distribution of each NE type differs greatly, we reserve only the first 500 articles for each NE type while sampling the training data randomly. Moreover, 0.5 percent of all articles in Chinese Wikipedia (3678 articles) are sampled from the article list at the interval of 20 articles and annotated as the test data.

Two independent annotators have been 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 the NE type of articles using richer information shown on the web page. We have calculated the Kappa coefficient (Carletta, 1996) for the annotated data. The Kappa score is 0.94, which indicates a good agreement between the annotators. These discrepancies have been then discussed by the two annotators, and the final results of this data are given after collective discussion.

Evaluation: We have adopted the widely used Precision, Recall, and *F*-score to measure the entity classification performance. The weighted average of each measure is used to evaluate the overall performance. The weighted average of precision is computed using the following method:

$$\text{Precision}_{\text{avg}} = \sum_i |c_i| \times \text{Precision}(c_i) \Big/ \sum_i |c_i|,$$

where $|c_i|$ (i=1, 2, 3, 4) is the number of the articles in NE type c_i, and Precision(c_i) is the precision of NE type c_i. The weighted averages of recall and *F*-score can be computed similarly.

Parameter: Using the training dataset, an SVM-based classifier, which is realized by the toolkit libSVM with a linear kernel, is trained to classify Wikipedia articles into NE types. For all classifiers, the dimension of the content feature vector is reduced to 2000 using IG, which can cover enough representative features in our experiments.

This training dataset is also used to obtain a set of optimal weight parameters (α_1, α_2, α_3, α_4). Firstly, we adopt only the method based on heuristic rules to classify Wikipedia articles, and it is considered to be a misclassification if there are equal maximum values in vector $\boldsymbol{C}_{\text{multi}}$ (including the zero-vector). Then, we define an objective function as the maximum overall precision in the training dataset, and loop through all combinations for parameters by a step length of 0.05. Finally, a set of optimal weight parameters is achieved (α_1=0.2, α_2=0.25, α_3=0.5, α_4=0.05).

We evaluate the performances of heuristic rules, the supervised NE classifier, and the combined method, respectively. The method of heuristic rules cannot classify the articles that contain equal maximum values in vector $\boldsymbol{C}_{\text{multi}}$, so we ignore these articles in the experiments. To validate the effectiveness of each feature set used by the supervised NE classifier, we train classifiers using the combination of the article content feature set and the other feature set, which are expressed as '+structured', '+category', and '+article title'. Furthermore, the classification model is trained based on all feature sets as the result of our supervised NE classifier. The results are shown in Table 1.

Table 1 Entity classification performance of heuristic rules, the supervised NE classifier, and the combined method

Method	Article number	Precision (%)	Recall (%)	*F*-score (%)
Heuristic rules	2391	95.36	95.46	95.40
Supervised NE classifier		87.67	87.14	87.28
Classifier (content)		81.09	80.21	80.37
Classifier (+structured)	3678	83.91	83.47	83.59
Classifier (+category)		85.25	84.75	84.79
Classifier (+article title)		85.69	85.51	85.52
Combined method	3678	91.31	90.56	**90.73**

The weighted average of each measure (precision, recall, and *F*-score) is used to evaluate the overall performance

The method of heuristic rules achieves the best performance in our experiments (Table 1), but there are a substantial number of articles that cannot be classified into NE types using this method. The improved performance is achieved by increasing each feature set to a supervised NE classifier, and these results show that they are useful in predicting the NE type of the article. Our combined method, which overcomes the limited coverage of heuristic rules and poor performance of the supervised NE classifier, achieves satisfactory classification results.

The classification results for each NE type are shown in Table 2. We can see that the entity classification of articles for the ORG type is more difficult than those for other NE types, which is similar to Chinese NER (Zhou *et al.*, 2006). Generally, the written forms of ORG article titles are various. For example, the length of the article title is uncertain (e.g., 'IBM' and 'China National Offshore Oil Corporation'), and the inner structure of the article title is complex, in which it contains elements of the person, location, upper-organization, and new words (e.g., for the article title '香港华仁书院' (Wah Yan College, Hong Kong), '香港' (Hong Kong) is the location element, '华仁' (Wah Yan) is the name element, and '书院' (College) is the keyword element).

Among all NE types, PER and LOC achieve the best performance (92.19% and 92.92% of *F*-score, respectively), but ORG is the worst (only 70.61% of *F*-score). However, there is a much lower percentage of ORG type (only 5.33%) in the test data, so this method still achieves excellent overall performance. Since there are many infrequently used entities in the wrongly classified articles for the ORG type, only part of classification errors transfer into automatically annotated corpora.

4.2 Evaluation of NER corpora

We evaluate our NER corpora from two aspects: the quality of automatically annotated corpora and the performance in practical applications. For the first aspect, we achieve the evaluation results by manual verification. For the second one, we design the experiments to compare the NER performances among different corpora.

4.2.1 Evaluation data

The following three sets of manually annotated NER data are chosen in our experiments: (1) MSRA NER corpus (SIGHAN, 2006); (2) the corpus of People's Daily tagged by Peking University (the training set contains data from January to March and the test set contains data in April); (3) a new corpus from online news that includes multiple domains, including economics, technology, politics, entertainment, sport, and military. The size of each corpus is listed in Table 3.

The Domains corpus has been created by manually annotating the text of 155 online news articles, downloaded from the websites of sina.cn and people.cn. For each text, we have annotated the text span and type of entity mentions and other text using special part-of-speeches ('nr' denotes PER type, 'nt' denotes ORG type, 'ns' denotes LOC type, and 'o' denotes other text), which have the same format as primitive MSRA test data. For example, the sentence '新华社吉隆坡１２月１６日电' (Kuala Lumpur, Dec. 16, Xinhua reported), is converted to annotated text '新华社/nt 吉隆坡/ns １２月１６日电/o'. The annotation process is finished by two independent annotators, and the final results of this data are given after collective discussion.

Table 2 Number of articles for each NE type in the test data and detailed results of combined method

NE type	Article number	Percentage (%)	Precision (%)	Recall (%)	*F*-score (%)
PER	632	17.18	86.95	98.10	92.19
ORG	196	5.33	61.92	82.14	70.61
LOC	1192	32.41	93.32	92.53	92.92
OTHER	1658	45.08	95.01	87.27	90.98

Table 3 Size of gold-standard corpora used for NER evaluation in our experiments

Parameter	MSRA		Peking		Domains
	Training	Testing	Training	Testing	Test
Token number	1 339 461	105 353	3 717 324	1 296 758	85 645
Sentence number	46 363	4364	61 619	21 389	2970
NE number	74 703	6181	151 278	59 690	4455

In the above corpora, three universal NE types (PER, ORG, and LOC) are tagged by special part-of-speeches (nr, nt, and ns). To maintain consistency with the criterion, we remove the original result of word segmentation in these corpora and realize Chinese word segmentation based on the same criterion. We evaluate the results using the CoNLL evaluation metric. CoNLL awards only exact phrasal matches, including the correct NE type and the text boundary.

4.2.2 Quality of corpora

The distribution of different NE types across different corpora is compared in Fig. 6. We can see that there are similar distributions between an automatically annotated corpus by Chinese Wikipedia (Wiki corpus) and the Peking corpus. For all corpora, the LOC type contains the largest percentage (approximately 50%). Although the density of the ORG type is increased using the defined criterion, the ORG type still has a lower percentage than the two gold-standard corpora (−11.6% for MSRA, −3.2% for Peking).

The automatically annotated corpora based on Chinese Wikipedia contain 2.3 million sentences (63 million tokens) that pass our criteria. To evaluate the quality of this corpus, 5000 sentences are selected from our enormous corpus randomly and annotated with NE labels manually as the Wiki gold-standard corpus (WG). We compare NE labels between the original corpus and WG, and the evaluation results are shown in Table 4.

There are two major types of errors for impacting the performance of automatic annotation. Firstly, although Wikipedia currently includes over 735 000 articles in Chinese alone, a large number of NEs (especially NEs of PER and ORG types) are not covered by this knowledge base. For example, the parents' names, which appear in Wikipedia article '大卫·贝克汉姆' (David Beckham), are not contained in Chinese Wikipedia. Secondly, the incorrect classification of non-entity Wikipedia articles leads to the errors of false positive. For example, the mentions '纪念品' (souvenir) are tagged with PER if this article is classified into the PER type.

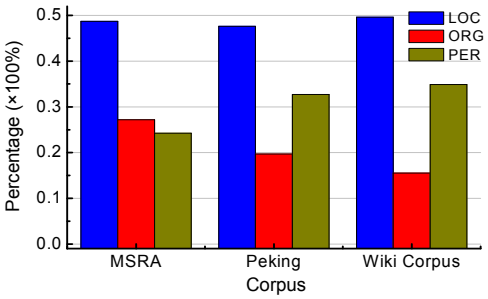

Fig. 6 NE distribution of different NE types across the gold-standard corpora and the Wiki corpus

Table 4 Evaluation results of automatically annotated corpora

NE type	Precision (%)	Recall (%)	*F*-score (%)
PER	78.89	81.83	80.33
ORG	81.38	84.86	83.08
LOC	78.26	60.38	68.17
Total	80.01	79.87	79.73

Balasuriya *et al.* (2009) evaluated the performance of automatic annotation based on English Wikipedia. Based on capitalization heuristics, they

retained sentences where it seemed that all NEs in the sentence had been tagged in the automatic process. In the process of Chinese automatic annotations, we adopt looser criteria because these heuristics cannot be adopted to identify a high-quality sentence. The overall performance of our automatic annotation is satisfactory to reduce the workload of manual annotations if these corpora are selected to annotate gold-standard corpora.

Moreover, we tag NEs in MSRA and Peking test data by finding implicit mentions. The automatically annotated results are treated as predicted annotations on MSRA and Peking test data, and the results are shown in Table 5. By comparing the results in Table 4 and Table 5, we can see that automatic annotations of Chinese Wikipedia achieve higher recall than those of MSRA and Peking test data, because there are more entity mentions corresponding to Wikipedia articles in the text of Chinese Wikipedia.

4.2.3 NER performance

For the predefined NE types, while the annotators want to build gold-standard corpora on the text from non-Wikipedia articles, the NER models are needed to realize the initial annotations for the text automatically, to reduce the workload of manual annotations. Moreover, these NER models trained by automatically annotated corpora reflect the quality of these corpora from another perspective.

To evaluate our corpora as NER training data, we use conditional random fields, which is one of the most common methods for Chinese NER. A customizable toolkit CRF++ is used to implement this algorithm. In our experiments, several common Chinese NER features are chosen, including Chinese word, part of speech, and their combinations. Moreover, 20 000 annotated sentences (about 540 000 tokens) are chosen because the performance cannot be improved with increasing corpora scales; however, the training time increases rapidly.

We conduct three groups of experiments with these corpora: (1) cross-evaluation among gold-standard corpora; (2) evaluation of automatically annotated corpora; (3) evaluation of combined corpora.

Three groups of evaluation results are shown in Table 6. The performance of the NER model decreases seriously when crossed training and test corpora are used (see G1 in Table 6). For example, the MSRA-trained NER model achieves an excellent *F*-score of 90.15% on the MSRA test corpus. However, this model does not perform well on other gold-standard test corpora (an *F*-score of −9.07% on the Peking test corpus; an *F*-score of −12.45% on the Domains corpus), and greater differences between the Domains corpus and the MSRA corpus on the genre and domain are an important reason that causes worse performance. Furthermore, we present the results on recognizing each entity type, and the best performance is on PER, followed by LOC, and the poorest performance is on ORG.

Then, three different corpus selection approaches are adopted to build Chinese Wikipedia corpora. The first one (denoted as CW-R) selects sentences randomly; the second one (denoted as CW-P) prefers to select sentences that contain more explicit mentions. Because most of mentions are tagged with corresponding outgoing links when they are first mentioned in Wikipedia articles, this approach prefers to select sentences in the first paragraph. The last one (denoted as CW-E) selects sentences based on our core article extending approach.

There are great differences between the CW-R corpus and gold-standard test corpus because the Wiki corpus contains automatically annotated sentences from a wider variety of domains. In our experiments, the NER model trained using the CW-R corpus achieves limited performance. The sentences in the CW-P corpus contain more explicit mentions that can precisely establish the mapping between

Table 5 Evaluation results of automatic annotation for MSRA and Peking test data

NE type	Precision (%)		Recall (%)		*F*-score (%)	
	MSRA	Peking	MSRA	Peking	MSRA	Peking
PER	83.93	76.44	65.92	42.84	73.84	54.91
ORG	66.91	49.07	54.24	51.25	59.92	50.14
LOC	79.41	77.79	67.87	86.51	73.19	81.92
Total	78.14	71.64	64.31	65.01	70.55	68.16

mention and the target article, so this approach achieves a better automatically annotated corpus than the CW-R approach. The CW-E approach can adopt the domains of the test corpus, and overcome the degradation of the domain transfer. The NER model trained by the CW-E corpus significantly outperforms other models by 2.90%–11.30% of *F*-score (see G2 in Table 6). By analyzing the results of each NE type in the G2 of Table 6, we can see that the performance of PER type decreases seriously on the CW-R corpus compared with gold-standard training data. A reasonable explanation is that, there are a large proportion of translated foreign names in Chinese Wikipedia, but more Chinese names in the test data. The model trained by the CW-E corpus can avoid this problem of inconsistency to a certain extent.

We also conduct an experiment with combined corpora consisting of gold-standard corpora and automatically annotated corpora (see G3 in Table 6). The combined corpora can play a complementary role by its basic feature recognition and domain transferring capabilities. The experimental results indicate that the combined corpora perform better than the single gold-standard corpus. Therefore, if there are gold-standard corpora for the predefined NE types, the combined corpora can be adopted to train the NER model to achieve better initial annotations for the text coming from different domains.

4.2.4 Evaluation of cross-domain

Only small improvements are achieved when combined corpora are used to evaluate the other gold-standard test data (+0.41% for MSRA+CW-E; +1.16% for Peking+CW-E). This result is caused by the high similarity between the MSRA corpus and the Peking corpus in the genre and domain. We compute the cosine similarity by building NE vectors on training and test data, and high similarities are achieved between MSRA and Peking corpora (85.9% between MSRA training data and Peking test data; 77.1% between Peking training data and MSRA test data).

To evaluate the NER performance in the transferred domain, we conduct an experiment on three special domains (economics, technology, and politics) from the domain corpus. For each test data from different domains, we have adopted the core article extending approach to build different training corpora (CW-E). The results are listed in Table 7.

The combination of the gold-standard corpus and our silver-standard corpus can automatically adapt the domain transferring of test data using our method (Table 7). In addition, the model trained with the combined corpus can effectively improve the NER performance.

Table 6 Evaluation results (*F*-score) of NER when training on different corpora

ID	Training corpus	*F*-score (%)											
		MSRA				Peking				Domains			
		PER	ORG	LOC	ALL	PER	ORG	LOC	ALL	PER	ORG	LOC	ALL
G1	MSRA	92.94	83.62	91.17	90.15	89.17	68.73	80.71	81.08	86.48	67.17	78.67	77.70
	Peking	93.12	59.44	76.61	79.22	94.49	84.01	91.38	91.18	85.66	50.00	76.51	72.53
G2	CW-R	54.57	62.65	74.95	66.36	55.91	47.71	80.25	67.20	53.16	63.95	68.41	62.38
	CW-P	66.91	50.39	76.12	68.46	69.37	49.25	81.40	71.75	72.88	50.39	69.57	64.55
	CW-E	72.57	61.38	80.89	74.45	70.96	55.12	82.84	74.65	78.15	66.39	75.56	73.68
G3	MSRA+CW-E	–	–	–	–	88.51	68.05	82.25	**81.49**	86.36	73.41	80.71	**80.02**
	Peking+CW-E	92.73	65.02	77.71	**80.38**	–	–	–	–	87.16	64.54	77.66	76.64

Table 7 Evaluation results (*F*-score) of NER for special cross-domain corpora

Test corpus	Number of tokens	*F*-score (%)				
		MSRA	Peking	CW-E	MSRA+CW-E	Peking+CW-E
Economics	24 346	76.16	70.11	70.14	**81.60**	75.45
Technology	18 488	76.34	70.68	72.03	**82.99**	79.19
Politics	20 883	81.26	77.58	75.47	**81.98**	78.54

5 Conclusions and future work

We have demonstrated a method of automatically building Chinese NER corpora using Chinese Wikipedia. To achieve better annotated corpora, the combination of the heuristic rules of multi-faceted information and the supervised NE classifier is adopted to accurately and comprehensively identify the NE type of Chinese Wikipedia articles. Moreover, we introduce a method for finding implicit mentions in Wikipedia article content and realize NE type identification of ambiguous mentions using the EL method. To obtain a better NER model as training corpora, we also present a tagged corpus selection approach based on core article extending, which can automatically adapt the domains of test data.

In our future work, we will focus on frequent entity mentions that may cause serious performance degradation. It will result in a large amount of errors in automatically annotated corpora if some frequent mentions (such as 'China') are classified into incorrect NE types. Besides, more strategies and resources are adopted to differentiate NEs from non-entities in Chinese Wikipedia.

References

Alotaibi, F., Lee, M., 2012. Mapping Arabic Wikipedia into the named entities taxonomy. Proc. 24th Int. Conf. on Computational Linguistics, p.43-52.

An, J., Lee, S., Lee, G.G., 2003. Automatic acquisition of named entity tagged corpus from World Wide Web. Proc. 41st Annual Meeting on Association for Computational Linguistics, p.165-168. [doi:10.3115/1075178.1075207]

Auer, S., Bizer, C., Kobilarov, G., *et al*., 2007. DBpedia: a nucleus for a Web of open data. *LNCS*, **4825**:722-735. [doi:10.1007/978-3-540-76298-0_52]

Balasuriya, D., Ringland, N., Nothman, J., *et al*., 2009. Named entity recognition in Wikipedia. Proc. Workshop on the People's Web Meets NLP, ACL-IJCNLP, p.10-18.

Bunescu, R., Paşca, M., 2006. Using encyclopedic knowledge for named entity disambiguation. Proc. 11th Conf. of the European Chapter of the Association for Computational Linguistics, p.9-16.

Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. *Comput. Ling.*, **22**(2):249-254.

Ciaramita, M., Altun, Y., 2005. Named-entity recognition in novel domains with external lexical knowledge. Proc. Human Language Technologies in Advances in Structured Learning for Text and Speech Processing Workshop, p.209-212.

Dakka, W., Cucerzan, S., 2008. Augmenting Wikipedia with named entity tags. Proc. Int. Joint Conf. on Natural Language Processing, p.545-552.

Darwish, K., 2013. Named entity recognition using cross-lingual resources: Arabic as an example. Proc. 51st Annual Meeting of the Association for Computational Linguistics, p.1558-1567.

Ehrmann, M., Turchi, M., 2010. Building multilingual named entity annotated corpora exploiting parallel corpora. Proc. Workshop on Annotation and Exploitation of Parallel Corpora, p.24-33.

Etzioni, O., Cafarella, M., Downey, D., *et al*., 2005. Unsupervised named-entity extraction from the Web: an experimental study. *Artif. Intell.*, **165**(1):91-134. [doi:10.1016/j.artint.2005.03.001]

Fu, R., Qin, B., Liu, T., 2011. Generating Chinese named entity data from a parallel corpus. Proc. 5th Int. Joint Conf. on Natural Language Processing, p.264-272.

Gabrilovich, E., Markovitch, S., 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. Proc. 20th Int. Joint Conf. on Artificial Intelligence, p.1606-1611.

Guo, H., Zhu, H., Guo, Z., *et al*., 2009. Domain adaptation with latent semantic association for named entity recognition. Proc. Human Language Technologies: the Annual Conf. of the North American Chapter of the ACL, p.281-289.

Higashinaka, R., Sadamitsu, K., Saito, K., *et al*., 2012. Creating an extended named entity dictionary from Wikipedia. Proc. 24th Int. Conf. on Computational Linguistics, p.1163-1178.

Ji, H., Grishman, R., Dang, H.T., 2011. Overview of the TAC2011 Knowledge Base Population Track. Proc. Text Analysis Conf.

Jiang, J., Zhai, C.X., 2006. Exploiting domain structure for named entity recognition. Proc. Main Conf. on Human Language Technology Conf. of the North American Chapter of the Association of Computational Linguistics, p.74-81. [doi:10.3115/1220835.1220845]

Jiang, J., Zhai, C.X., 2007. A two-stage approach to domain adaptation for statistical classifiers. Proc. 16th ACM Conf. on Information and Knowledge Management, p.401-410. [doi:10.1145/1321440.1321498]

Liao, W., Veeramachaneni, S., 2009. A simple semi-supervised algorithm for named entity recognition. Proc. NAACL HLT Workshop on Semi-Supervised Learning for Natural Language Processing, p.58-65.

Liu, H., Chen, Y., 2010. Computing semantic relatedness between named entities using Wikipedia. Proc. Int. Conf. on Artificial Intelligence and Computational Intelligence, p.388-392. [doi:10.1109/AICI.2010.88]

Liu, X., Zhang, S., Wei, F., *et al*., 2011. Recognizing named entities in Tweets. Proc. 49th Annual Meeting of the Association for Computational Linguistics, p.359-367.

Medelyan, O., Milne, D., Legg, C., *et al*., 2009. Mining meaning from Wikipedia. *Int. J. Human-Comput. Stud.*, **67**(9):716-754. [doi:10.1016/j.ijhcs.2009.05.004]

Mika, P., Ciaramita, M., Zaragoza, H., *et al*., 2008. Learning to

tag and tagging to learn: a case study on Wikipedia. *IEEE Intell. Syst.*, **23**(5):26-33. [doi:10.1109/MIS.2008.85]

Nadeau, D., Turney, P.D., Matwin, S., 2006. Unsupervised named entity recognition: generating gazetteers and resolving ambiguity. *LNCS*, **4013**:266-277. [doi:10.1007/11766247_23]

Nastase, V., Strube, M., 2013. Transforming Wikipedia into a large scale multilingual concept network. *Artif. Intell.*, **194**:62-85. [doi:10.1016/j.artint.2012.06.008]

Nemeskey, D.M., Simon, E., 2012. Automatically generated NE tagged corpora for English and Hungarian. Proc. 4th Named Entity Workshop, p.38-46.

Ni, Y., Zhang, L., Qiu, Z., *et al*., 2010. Enhancing the open-domain classification of named entity using linked open data. Proc. 9th Int. Semantic Web Conf., p.566-581.

Nothman, J., Curran, J.R., Murphy, T., 2008. Transforming Wikipedia into named entity training data. Proc. Australian Language Technology Workshop, p.124-132.

Nothman, J., Ringland, N., Radford, W., *et al*., 2013. Learning multilingual named entity recognition from Wikipedia. *Artif. Intell.*, **194**:151-175. [doi:10.1016/j.artint.2012.03.006]

Ratinov, L., Roth, D., 2009. Design challenges and misconceptions in named entity recognition. Proc. 13th Conf. on Computational Natural Language Learning, p.147-155. [doi:10.3115/1596374.1596399]

Richman, A.E., Schone, P., 2008. Mining Wiki resources for multilingual named entity recognition. Proc. 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, p.1-9.

Toral, A., Ferrández, S., Monachini, M., *et al*., 2012. Web 2.0, language resources and standards to automatically build a multilingual named entity lexicon. *Lang. Res. Eval.*, **46**(3):383-419. [doi:10.1007/s10579-011-9148-x]

Zesch, T., Müller, C., Gurevych, I., 2008. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. Proc. Conf. on Language Resources and Evaluation, p.1646-1651.

Zhang, W., Sun, L., Zhang, X., 2012. A entity relation extraction method based on Wikipedia and pattern clustering. *J. Chin. Inform. Process.*, **26**(2):75-81 (in Chinese).

Zhou, J., Dai, X., Yin, C., *et al*., 2006. Automatic recognition of Chinese organization name based on cascaded conditional random fields. *Acta Electron. Sin.*, **34**(5):804-809 (in Chinese).

Appendix: Main patterns for entity classification

Table A1 Main patterns of categories for entity classification

NE type	Patterns of category
PER	Common concept: [+]人[物] (all kinds of people, such as 'people from xx' and 'poets'), [+]学家 (all kinds of experts, such as 'scientists' and 'biologists') Occupation or status: [*]演员 (actors), [*]歌手 (singer), [*]作家 (writer), etc. Special time: [前]<TIME>出生 (time of birth), [前]<TIME>逝世 (time of death)
ORG	Related companies: [+]公司 (companies), [+]机构 (organizations), [+]企业 (enterprises) Political unions: [+]政党 (political parties), [+]部门 (departments) Other organizations: [+]大学 (universities), [+]学校 (schools), [+]俱乐部 (clubs), [*]出版社 (publishing houses), [+]团体 (groups), etc.
LOC	Regions: [+]国家 (countries), [+]城市 (cities), [+]城镇 (towns) Natural locations: [+]河流 (rivers), [+]山 (mountains), [*]山脉 (mountain ranges), [+]地理 (geography), etc. Public places: [+]广场 (squares), [+]博物馆 (museums), [+]车站 (stations), etc.
DAB	[*]消歧[*] (disambiguation)
OTHER	[*]植物 (plants), [*]动物 (animals), [*]游戏 (games), [*]电影 (films), etc.

The meaning of expressions is interpreted in Section 3.1.1

Table A2 Main patterns of article titles for entity classification

NE type	Patterns of surface name
PER	Parenthesized expression: [*]演员 (actor), [*]歌手 (singer), [*]运动员 (athlete), [*]作家 (author), etc.
ORG	Last common noun: [+]集团 (group), [+]银行 (bank), [+]交易所 (exchange), [+]大学 (university), [+]出版社 (press agency), [+]公司 (company), [+]电视台 (television station), etc.
LOC	Last common noun: [+]国道 (the national road), [+]火车站 (railway station), [+]植物园 (botanical gardens), [+]大桥 (bridge), [+]石窟 (grottoes), etc.
OTHER	Last common noun: [*]列表 (list of …), [*]事件 (event), [*]奖 (prize), [*]车 (vehicle), etc. Parenthesized expression: [*]电影 (film), [*]电视剧 (TV play), [*]游戏 (game), [*]歌曲 (song), [*]动画 (animated cartoon), etc.