Towards an Enhanced and Adaptable Ontology by Distilling and Assembling Online Encyclopedias*

Shan Jiang
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
sjiang18@illinois.edu

Lidong Bing
School of CES, Shanghai University
Department of SEEM, The Chinese University of Hong Kong
ldbing@se.cuhk.edu.hk

Yan Zhang†
Department of Machine Intelligence
Peking University
Beijing 100871, China
zhy@cis.pku.edu.cn

ABSTRACT

In this paper, we investigate the problem of making better use of semantic knowledge obtained from different encyclopedia sources. We propose a framework to integrate different encyclopedias and reorganize the information. We also utilize Learning to Rank models to distill out more functional knowledge from the encyclopedic information and then align the knowledge with a WordNet-like ontology. Finally as a demonstration, a Chinese semantic knowledge repository named JNet is constructed based on this framework. Experiments show that the proposed methods work well and the three steps reinforce each other towards a more powerful ontology.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Dictionaries, Thesauruses*

Keywords

JNET, CCD, Encyclopedia, Ontology

1. INTRODUCTION

In the era of information explosion, the deluge of knowledge causes an unremitting emergence of neologisms. As a result, the vocabulary knowledge needed to understand a language dwarfs the coverage of any existing manual lexical knowledge base. Therefore, an automatically built lexicon with wider coverage is of great help for interpreting semantic information. The knowledge from the folk wisdom amassed in the online encyclopedias is a rewarding resource owing to the constant contribution of the numerous users. Web users are likely to introduce a neologism to an online encyclopedia as soon as it comes into vogue. In contrast, neologisms will be added into an expert-edited lexicon in the next edition several months or even several years later.

This problem is even more urgent in Chinese. Different from English and many other languages, the Chinese words are not separated from each other by the space symbol in writing. As a result, term segmentation is the primary step in Chinese language processing. If the semantic repository employed in this step is not able to recognize the words and separate them correctly, the semantic information of the sentences will be misinterpreted or even lost. Therefore, a powerful semantic repository is of significant importance in Chinese language processing. In this paper, we mainly focus on Chinese ontology enhancement.

The online encyclopedia, albeit rich in knowledge storage, is casually organized. Besides, overlaps exist between different encyclopedias and even data obtained from the same source needs to be reorganized. In encyclopedias like English Wikipedia, most ambiguous words are separately explicated in different articles. Nevertheless, in most Chinese online encyclopedias which are not well developed, words are sometimes organized in terms of word forms and different meanings of a word are exhibited in the same page. Besides, alternative names for the same concepts are scattered in different articles. To merge online encyclopedias into a semantic lexicon using concepts as basic elements, homograph disambiguation and synonym detection are prerequisites.

Although a broad vocabulary knowledge is one of the main aims to build an enriched ontology, it is in effect sometimes unnecessary to retain all the information obtained from the online encyclopedias. Because of the collaborative way of editing, the quality of some articles in online encyclopedias is rather low. Noises will be brought in by spam articles and impair the quality of the enriched ontology. Moreover, some supplementary knowledge, on one hand is precious treasure, on the other hand could be a big burden in practical applications sometimes. An ontology encapsulating all the entries of an online encyclopedia will include millions of words, some of which are quite rarely used. In comparison, expert-edited lexicons such as WordNet usually contain tens of thousands of words. For applications which utilize the ontology as a basic tool, the space and time consumption will be increased by a large scale due to the huge size of the ontology. We believe that the supply of the supplementary knowledge should be abundant but not superfluous.

In this paper, we integrate two online Chinese encyclopedias with a WordNet-like semantic lexicon of contemporary Chinese named Chinese Concept Dictionary (CCD) and construct an enriched ontology. We firstly crawl articles from the online ency-

*Supported by NSFC with Grant No. 61073081 and 973 Program with Grant No. 2014CB340405
†Corresponding author

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505597 ...$15.00.

clopedias, and then the encyclopedia articles are reorganized and grouped into synsets in the next step. After that, the synsets are ranked according to several criteria. Then different versions of the enriched ontologies are constructed by adding different proportions of the highly-ranked synsets from the encyclopedias to the expert-edited lexicon.

2. ENCYCLOPEDIA INTEGRATION

The main goal of encyclopedia integration is to reorganize the encyclopedic information in terms of word meanings. Entities with the same name but referring to different meanings should be discriminated and entities which are synonymous but have different names should be assembled.

2.1 Background

Encyclopedia entities are the basic units of an encyclopedia and are presented by encyclopedia articles:

Encyclopedia Entity: *An encyclopedia entity is composed of a six-tuple* $\langle$`name, Edt, Intro, Cites, Rel, Ctg`$\rangle$. *Here* `name` *is the name of the entity, initially composed of its article title. After integration,* `name` *could be a synset containing several alternative names of the entity.* `Edt` *refers to the edit information and is composed of sets of two-tuples* `<Edtor, Num>` *which means the editor* `Edtor` *has edited the article for* `Num` *times. And* `Intro` *is the introduction of the entity, namely the text of the article body.* `Cites` *denotes the set of cites which appear in the article and link to other articles.* `Rel` *indicates the set of related words and* `Ctg` *denotes the set of categories the entity belongs to.*

Concepts are fundamental elements of a WordNet-like ontology. Words of the same meaning are packed into the synonym set of a concept and a semantic network is constructed by linking concepts with various semantic relations. In this paper, we mainly focus on hyponymy relation.

Ontology Concept: *An ontology concept* ς *in a WordNet-like ontology is denoted as a four-tuple* $\langle$`Syn, Def, Hyper, Hypo`$\rangle$, *where* `Syn` *is the synonym set,* `Def` *is the definition of the concept,* `Hyper` *and* `Hypo` *denote the hypernym and the hyponym set of the concept respectively.*

2.2 Encyclopedia Entity Discrimination

The encyclopedia entity discrimination (EED) aims at discriminating homographic entities. We mainly use the encyclopedia articles to tackle the task. If two articles have very similar content, they are likely to refer to the same meaning for the following reasons. i) The introduction of the article should contain some basic information (e.g., when introducing the Apple company, it is necessary to point out that it is an American corporation and sells consumer electronics, computers, etc.). ii) Due to the open way of editing and lack of strict copyright protection, editors sometimes borrow sentences or even paragraphs from other encyclopedias during editing.

However, similarity measurement based on bag-of-words or n-gram model may misdirect the identification results when the divergent part is larger than the consistent part between the articles. A solution to this problem is to measure the similarity between two articles based on the common contiguous sequence (CCS) in them:

Definition 1: ***CCS*** - *Considering two text fragments represented as* $t_1 = \langle w_{11}, w_{12}, ..., w_{1m}\rangle$ *and* $t_2 = \langle w_{21}, w_{22}, ..., w_{2n}\rangle$, *where* w_{ij} *denotes the* j*-th word in* t_i (i=1 *or* 2). *Sequence* $\langle w_{1,p}, w_{1,p+1}, ..., w_{1,p+l}\rangle$ $(1 \le p \le m-l$ *and* $l \ge 0)$ *is a CCS of* t_1 *and* t_2 *if there exists* q $(1 \le q \le n-l)$, *such that* $w_{1,p} = w_{2,q}$, $w_{1,p+1} = w_{2,q+1}$, ..., $w_{1,p+l} = w_{2,q+l}$.

Definition 2: ***closed CCS*** - *Following the notations used in Definition 1,* $CCS_1 = \langle w_{1,p}, w_{1,p+1}, ..., w_{1,p+l}\rangle$ *will be a closed CCS of* t_1 *and* t_2 *if there does not exist any other* $CCS_2 = \langle w_{1,o}, w_{1,o+1}, ..., w_{1,o+r}\rangle$ $(r > l)$ *in which we can find an integer* k $(0 \le k \le r-l)$ *making* $w_{1,p} = w_{2,o+k}$, $w_{1,p+1} = w_{2,o+k+1}$, ..., $w_{1,p+l} = w_{2,o+k+l}$.

Given two articles, we make use of closed CCS to measure the similarity between them. To eliminate very short CCSs, a length threshold θ_L is set to select the qualified closed CCSs. The lengths of the qualified closed CCSs are denoted as l_1, l_2, ..., l_m.We define a function $f(l_1, ..., l_m)$ to measure the similarity between two texts, and propose some constraints which should be satisfied:

$a)$ $\forall 1 \le i \le m,\ \ f(l_1, ..., l_m) > f(l_i)$;
$b)$ $\forall m > 1,\ \ f(l_1, ..., l_m) < f(\sum_{i=1}^{m} l_i)$;
$c)$ $(\sum_{i=1}^{m} l_i \ge \sum_{j=1}^{n} l'_j \wedge m \le n) \Rightarrow f(\sum_{i=1}^{m} l_i) \ge f(\sum_{j=1}^{n} l'_j)$.

Constraint $a)$ is designed to guarantee that texts with more closed CCSs are more similar. Constraint $b)$ is based on the insight that a single contiguous sequence is a more convincing evidence to reveal high similarity than the discretely distributed ones with the same length in summation. Constraint $c)$ ensures that CCSs with stronger contiguity or larger length in summation contribute more to the similarity. The following formula can be proved fulfilling all the three constraints:

$$f(l_1, ..., l_m) = ln[m + e^{1+\sum_{i=1}^{m}(l_i-1)} - 1]. \quad (1)$$

If entity e_1 and e_2 have the same name and the value of f function of $e_1.Intro$ and $e_2.Intro$ is larger than a threshold θ_f, e_1 and e_2 will be identified as an identical entity. If they are not recognized as being identical by the CCS-based method, related words and category information will be utilized to give guidance for further decision. Different from the text of the detailed introduction, related words and categories are more elaborately selected and more semantically coherent with the entity. If the proportion of the overlap to the total number of categories and related words is higher than a threshold θ_{rc}, e_1 and e_2 will be treated as an identical entity.

2.3 Synonym Detection

To further leverage the semantic information, it is necessary to assemble entities which are synonymous but with different names together. Instead of randomly comparing two entities and checking whether they refer to the same meaning, we automatically extract synonyms of an entity and collect entities having overlaps in their synonym set as candidates. Primarily, redirection information can provide important hints. If two entities have a redirection link between them, they are confirmed to be synonymous. A supplementary step is to extract alternative names from the introduction of the entities by using patterns such as "*A is known as B*". Then entities sharing a common alternative name are further checked by the method used for EED.

The procedure of the encyclopedia integration is summarized as follows. There are three major steps in the main routine, namely, alternative name extraction, redirected entity merging and entity discrimination. The first two steps are straightforward as discussed above. In the third step, the discrimination operation is performed for each pair of entities if they have not been disambiguated in any encyclopedia yet and have a common name.

3. ENCYCLOPEDIC INFORMATION DISTILLING

In general, encyclopedic information can be viewed as a rich knowledge resource. For some applications such as the search engine, bringing in all the knowledge is helpful. However, for some other applications which is sensitive to the size of the ontology, it is necessary to keep balance between the vocabulary coverage and the practicability of the ontology. To make the new ontology

more adaptive, we distill the encyclopedic information by casting the problem as an entity ranking problem. Entities ranked higher will be more likely to be functional knowledge and lower-ranked entities tend to be more narrowly-used or noise.

3.1 Ranking Encyclopedia Entities

Two types of Learning to Rank model are employed to solve the entity ranking problem in this paper. One is pairwise approach and the other is pointwise approach. SVM-Rank is employed to conduct the pairwise approach in our framework. It solves the ranking problem with Support Vector Machine (SVM) approach [10]. Partial orders should be given to each pair of the training examples and the ranking order of the rest data will be automatically determined by SVM-Rank. Different from pairwise approach, pointwise approach reduces the ranking problem to a regression or classification problem on single elements. Linear Regression based on formulae proposed by [26, 25] is carried out to implement the pointwise approach. It estimates the target variable as the value of an affine function of one or more explanatory variables.

Entities are represented by vectors in a space which takes features reflecting relevant characteristics of the entities as dimensions. Given manually rated examples, the rest entities will be automatically ranked by the trained models from these examples.

3.2 Features for Entity Ranking

We employ five features to represent entities in both methods. Among them, web popularity (WP) and TrustRank (TR) are features to capture the importance of the entity. Edit quality (EQ), edit times (ET) and article length (AL) are intended to reveal the quality of the article. Web popularity is measured by the number of results returned by a search engine when setting the entity name as a query. Edit times is how many times the article is edited and article length equals to the number of words in the introduction. The collection of the above three features is straightforward and the details are omitted due to space limitation. More details about TrustRank and edit quality are given as follows.

3.2.1 TrustRank

We employ the TrustRank model [5] to separate the valuable articles introducing well-known entities from the less important ones introducing little-known entities. Primarily, inverse PageRank of each article is calculated to select seed articles and a certain number, say K, of articles with high inverse PageRank scores are selected as seeds and manually marked as:

$$\mathbf{d}(a_i) = \begin{cases} 1, & if\ a_i\ is\ valuable; \\ 0, & otherwise. \end{cases} \quad (2)$$

a_i ($1 \leq i \leq K$) is a selected article and $\mathbf{d}$ is an N-dimension vector. For other articles which are not selected as seeds, their corresponding values in vector $\mathbf{d}$ will be set to be β_d ($0 < \beta_d < 1$). The vector $\mathbf{d}$ is normalized as $\mathbf{d} = \mathbf{d}/|\mathbf{d}|$ and the TrustRank vector $\mathbf{t}$ is initialized as $\mathbf{d}$ in the next step.

The TrustRank scores are finally computed by iterating the following formula:

$$\mathbf{t} = (1-\alpha) \cdot \mathbf{T} \cdot \mathbf{t} + \alpha \mathbf{d}, \quad (3)$$

where α is a dampening factor and $\mathbf{T}$ is a transition matrix. $\mathbf{T}_{ij} = t_{ij}/B_j$, where t_{ij} is the number of links from the j-th article to the i-th article, and B_j is the number of outlinks of the j-th article.

3.2.2 Edit quality

In general, the more experienced an editor is, the more probable it is that the edited articles are of good quality. And the more high-quality articles an editor have edited, the more skilled the editor tends to be. Taking advantage of this mutual reinforcement relationship, we can involve the HITS algorithm [13] to estimate the quality of articles. A formal definition is given as follows:

$$\mathbf{q} = \mathbf{M} \cdot \mathbf{a}\,; \quad (4)$$

$$\mathbf{a} = \mathbf{M}^T \cdot \mathbf{q}\,. \quad (5)$$

$\mathbf{q}_i$ is the quality of the i-th article and $\mathbf{a}_j$ is the editing ability of the j-th editor. $\mathbf{M}_{ij}$ means how many times the j-th editor has edited the i-th article. After some times of iteration, the vector $\mathbf{q}$ and $\mathbf{a}$ will converge. Then, vector $\mathbf{q}$ gives the edit quality of the articles.

4. ALIGNING ENCYCLOPEDIC KNOWLEDGE WITH ONTOLOGY

The online encyclopedias are rich in semantic relations, but the relations are neither labeled nor arranged. In WordNet, words are packed into synsets and various relations are hold between them, among which hyponymy is one of the most universal relations. Based on hyponymy relation between concepts, a tree structure of the ontology can be built. To make better use of the existing structure of a WordNet-like ontology, we reorganize the encyclopedic information by aligning the two kinds of semantic knowledge resources through finding the hypernym-style categories of entities.

Generally, categories are sometimes likely to be the hypernym of the entity because "*A belongs to category B*" implies that *A* is a kind of *B*. However, some categories are not limited to the hyponymy relation but tend to be a kind of tag information.

A special section named "related words" is contained by most Chinese encyclopedias. These words are elaborately selected by editors and are closely related to the entity. For example, related words for individuals are usually people who work in the similar fields or have personal relations with the target entity. We make use of this special resource to find hypernym-style categories and propose an algorithm based on several observations: i) hypernym-style categories tend to be in the same branch in the hyponymy tree because they are represented as ancestor nodes of the target entity, while the tag-style categories are likely to be in other branches which do not contain the target entity as a family member; ii) the related words of the target entity are more semantically related to the hypernym-style categories than the tag-style categories because they tend to be located in the neighborhood of the target entity; iii) reasonable hypernym-style categories are more likely to be recommended in different encyclopedias than tag-style categories.

Given an entity e, we first find the lowest super-ordinate of its categories in the hyponymy tree and denote it as lso. Then the semantic relatedness between the related words and every branch whose root is one of the child nodes of lso is calculated as:

$$SR(r, e.Rel) = \begin{cases} \frac{\sum_{\varsigma \in \Phi} \sum_{w \in e.Rel} sim(w,\varsigma) \cdot log(N_\varsigma + 1)}{|\Phi|}, & if(|\Phi| \neq 0); \\ 0, & else. \end{cases} \quad (6)$$

where r is the root of the branch and Φ is set of nodes which are in the branch (including r and its descendants) and contain any element of the category set of e (i.e., $e.Ctg$). $e.Rel$ is the set of related words of e and N_ς is how many times category ς appears in different encyclopedias. We use $log(N_\varsigma + 1)$ instead of N_ς to confine the influence of N_ς. Since $sim(w, \varsigma)$ (it will be introduced later) is less than 1, the SR value will be dominated by N_ς if N_ς is used directly. Hence the log form is taken and the logarithmic base is set to be the number of encyclopedias used. $sim(w, \varsigma)$ is the semantic similarity between w and ς and is defined as:

$$sim(w, \varsigma) = \begin{cases} 1, & if\ (w = \varsigma); \\ \frac{log \frac{len(w,\varsigma)}{depth(w)+depth(\varsigma)}}{log \frac{1}{2(\max\limits_{c \in \Omega} depth(c)+1)}}, & if((w \neq \varsigma) \wedge (w \in \Omega \wedge \varsigma \in \Omega)); \\ 0, & if((w \neq \varsigma) \wedge (w \notin \Omega \vee \varsigma \notin \Omega)). \end{cases} \quad (7)$$

where Ω is the hyponymy tree. $len(w, \varsigma)$ is the length of the shortest path between w and ς in Ω. $depth(c)$ is the depth of c and equals to the length of the path from c to the root of Ω. The intuition for this formula is: 1) the shorter the path between the concepts is, the more related the concepts are; 2) concepts with a deeper depth seem to be more closely related than the shallower ones which have the same length of shortest path between them [9].

The sub-branch with the highest SR score is selected and the search scope is narrowed to this branch. The above procedure is repeated until the branch contains no element of the category set except the root.

In this way, encyclopedia entities will be mapped to the WordNet-like ontology and a new ontology named JNet is constructed . JNet is generated from the entities integrated from more than one encyclopedia and different versions of JNet can be built by adding different proportion of highly ranked entities. The framework for construction of JNet is summarized in Algorithm 1.

Algorithm 1: Framework of construction of JNet

1 $E \leftarrow Integrate\ encyclopedias(E_1, E_2, ...)$
2 calculate the values of WP, TR, EQ, ET, AL for each entity in E and represent the entities in the feature space
3 $RankOrder(E)$ = Ranking order of E given by Learning to Rank model
4 set a proportion $p\%$
5 $top(E, p\%)$ = the top $p\%$ of the entities based on $RankOrder(E)$
6 JNet $\leftarrow Merge\ ontologies(top(E, p\%)$, WordNet-like Ontology $O)$

5. EXPERIMENTS

5.1 Data and Resource

In our experiments, we build several different versions of JNet in Chinese. Chinese online encyclopedias have a booming development in recent years. Some of them have comparable scale of English Wikipedia and are far beyond Chinese Wikipedia. Two largest Chinese online encyclopedias are used, namely Baidu Baike[1] and Hudong Baike[2]. The data was crawled from Nov. 15th 2011 to Nov. 30th 2011, containing 3,819,124 and 1,190,528 articles of Baidu Baike and Hudong Baike respectively. In the ontology merging stage, the WordNet-like ontology employed is Chinese Concept Dictionary (CCD) [28, 14]. We use the version of CCD updated in June 2009, which is the latest version. It contains 99,642 concepts with 142,913 different word forms.

5.2 Evaluation of Data Integration

5.2.1 Evaluation criteria for EED

As mentioned in Section 2.2, the target of encyclopedia entity discrimination (EED) is to distinguish two classes correctly, namely entities with the same name but referring to different meanings (homographs) and entities having the same name as well as the same meaning (identical entity). Basically, we can use F-measure to evaluate the results. There are two types of F-measure score which are widely used, i.e., Macro-average F-measure score (F^{ma}) and Micro-average F-measure score (F^{mi}).

Macro-average F-measure score: Primarily, the F-measure for class c_i is calculated as $F_i = \frac{2P_iR_i}{P_i+R_i}$. P_i and R_i are the precision and recall for each class c_i respectively. Then the macro-average F-measure is $F^{ma} = \frac{\sum_i F_i}{|C|}$.

Micro-average F-measure score: The global precision and recall are defined as: P=$\frac{\sum_i TP_i}{\sum_i (TP_i+FP_i)}$, R=$\frac{\sum_i TP_i}{\sum_i (TP_i+FN_i)}$, where TP_i, FP_i and FN_i are true positive, false positive and false negative number for class c_i. The Micro-average F-measure is F^{mi}=$\frac{2PR}{P+R}$.

[1]http://baike.baidu.com/
[2]http://www.hudong.com/

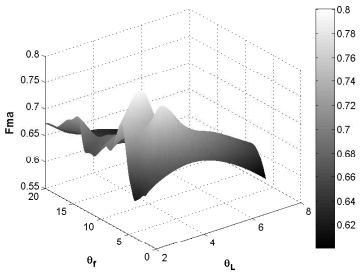


Figure 1: F^{ma} value of EED with the combination of different θ_L and θ_f ($\theta_{rc} = 0.17$).

Given that identical entities outnumber homographs by a wide margin in practice, we adopt the Macro-average F-measure score (F^{ma}) to evaluate the performance of EED.

5.2.2 Experiment preparation

For encyclopedia entity discrimination, the three parameters mentioned in Section 2, namely θ_L, θ_f and θ_{rc}, are determined by optimizing the performance on a random sample of 500 article-pairs. Each pair is composed of two articles with the same title. Generally, the performance of θ_{rc} is comparatively stable. The effects of θ_L and θ_f are much greater than that of θ_{rc} and the results with combination of different θ_L and θ_f are shown in Figure 1. The experimental results show that the best performance is achieved when θ_L, θ_f and θ_{rc} are set to be 3, 7 and 0.17 respectively.

For synonym detection, the patterns utilized for synonym extraction are shown in Table 1.

Table 1: Patterns for synonym extraction.

Chinese patterns	English translation
*A*也\又\亦(被)叫\称(做\为)*B*,	*A* is also named *B*,
*A*被称为(是)*B*,	*A* is known as *B*,
*A*又名(叫)(做)*B*,	*A* is also called *B*,
A(的)别名(是\为)*B*	The alias of *A* is *B*

5.2.3 Results and discussion

In total, 121,602 entity pairs from the intersection of the two encyclopedias are identified as identical entities and the rest 4,855 words are marked as homographs. To check whether the selected parameters also work for the rest part of the data, we randomly sample another 500 article-pairs. A manual check finds 8 of them are homographs and the remaining 492 pairs are identical entities. Our method classifies 21 of them to be homographs and the rest 479 of them as identical entities, among which 7 of the homographs and 478 of the identical entities are correctly classified, leading to an F^{ma} score of 0.734.

For synonym detection, we find 77,986 redirection links in total. To ensure that the automatic extraction of alternative names brings as less noise as possible, the patterns are used very strictly. Finally, 446 synonym sets are automatically extracted from the articles and 423 of them are correct, which leads to a precision of 94.84%.

5.3 Evaluation of Ontology Merging

The quality of automatic merging is affected by the quality of the articles. An entity may be mapped to an improper position in JNet if most of its categories given by the editors are not closely related to it. To evaluate the performance of the mapping algorithm for entities at different ranking positions, the integrated entities are divided into four bins and the mapping results of an example set in each bin are manually checked.

The first bin (Bin1) is composed of the top 20% entities and the second bin (Bin2) includes entities ranked from 20% to 50%. Entities ranked from 50% to 80% constitute the third bin (Bin3) and the bottom 20% entities are put into the fourth bin (Bin4). 150 mapping cases are randomly sampled from each bin. Based on the two different ranking orders given by the Learning to Rank methods, we totally sample 1,200 cases.

The method proposed by Yamada and Torisawa [27] is also implemented. Since neither Baidu Baike nor Hudong Baike provides semi-structured article data as MediaWiki [3] does in Wikipedia system, the categories are used as the source for hypernym acquisition. Among the hypernym candidates, the one which generates the highest SVM score is regarded as the final winner. Besides, stemming is not very useful for dealing with Chinese encyclopedias so suffix is not used as a feature. Yamada's method achieve a precision of 73.25%. Table 2 shows the precision of mapping for the samples from each bin.

Table 2: Precision of mapping in each bin.

		Bin1	Bin2	Bin3	Bin4	AVG
Proposed Method	SVM	83.33	74.00	86.67	79.33	80.25
	LR	84.67	78.00	80.00	76.00	
Yamada's Method	SVM	76.00	73.33	70.00	66.67	73.25
	LR	85.33	74.00	76.00	64.67	

Highly-ranked entities are mapped more accurately because they are generally better edited and the category information can provide better guidance for the mapping. For some entities which are not rated highly, the related words and category information are sometimes quite limited. We manually check some uncorrect cases in Bin4 and find that the reason for misjudgement of the hypernym-style category is that all the categories are tag-style.

Yamada's method works very well in Japanese Wikipedia while it is not so effective in the Chinese encyclopedias used in this paper. One main reason is that the tool for hypernym acquisition [21] used in Yamada's method is not applied to the Chinese encyclopedias used in this paper since they do not provide the required semi-structured data of the articles. As reported by Yamada and Torisawa [27], the tool can achieve a precision of 90%. The inapplicability of this powerful tool may be responsible for the underperformance of Yamada's method in Chinese encyclopedias.

As discussed above, the ranking of entities affects the mapping accuracy. To evaluate the ranking results more directly, the matching degree of the ranking given by the Leaning to Rank methods and the ranking given by human is investigated. We invite volunteers to give manual ratings of a random sample of 200 examples from the integrated entities. Each entity is given a numerical score out of 100 and the criteria are the same as those in the scoring of training samples. SVM-Rank achieves an $NDCG$ value of 0.872 and Linear Regression achieves a value of 0.873.

5.4 Application of JNet in Text Mining

Besides the utility in Natural Language Processing (NLP), semantic knowledge repositories are also widely used in other fields like information retrieval and text mining. Considering that text mining tasks can involve huge amount of data which will challenge the knowledge coverage of an ontology, we employ different versions of JNet in some text mining tasks to see whether the refined ontologies can retain valuable information with a smaller size.

5.4.1 Experimental setting

We construct four different versions of JNet. The proportion $p\%$ in Algorithm 1 is set to be 20%, 50%, 80% and 100%, and the corresponding JNets constructed are denoted as JNet1, JNet2, JNet3 and JNet4 respectively.

Two ontology-based feature selection methods are adopted to conduct the experiment, namely, the method proposed by Hotho et al. [7] and TCRL [9]. Hotho's method reflects a synset with a fix number of levels of its hypernyms as one dimension. TCRL restructures the feature space more dynamically. Leaf nodes will be merged to their parent nodes in the hyponymy tree by TCRL recursively, on condition that the parent nodes have a stronger ability of representing the documents.

In our experiments, we use a document collection provided by Sohu, one of the most famous Internet corporations in China. It contains 1,165,452 news documents downloaded from "Sohu News"[4] in Dec. 2009. The documents are manually categorized into 13 pre-defined classes. We use five classes and randomly select 1000 documents from each class. We conduct experiments on text classification and clustering with SVM (LibSVM [3] is employed) and K-Means algorithms (Cluto [12] is employed) respectively. For clustering, purity [29] is employed as the evaluation criterion. For classification, Macro-average F-measure score (F^{ma}) and Micro-average F-measure score (F^{mi}) are calculated respectively.

5.4.2 Results and discussion

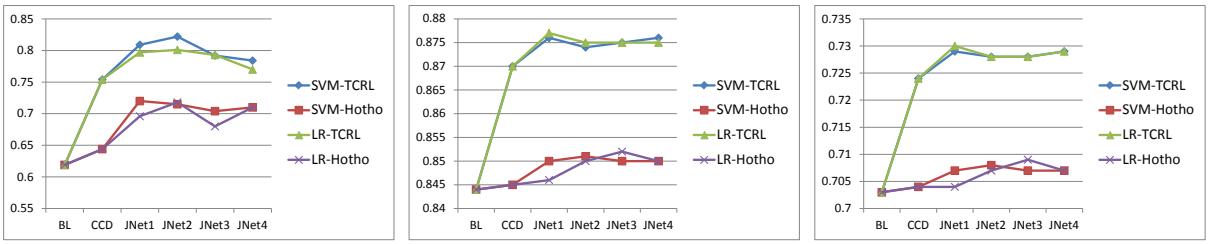
The results are shown in Figure 2. "BL" denotes the baseline which uses the original data without feature selection. "SVM-Hotho" means the results are obtained by using feature selection method proposed by Hotho et al. and employing JNets with entity distilling based on SVM-Rank. "LR-TCRL" uses TCRL and Linear Regression, and so forth. Both Hotho's method and TCRL outperform the baseline, which indicates that borrowing semantic information from the ontology will benefit text mining. All versions of JNet perform better than CCD because JNet imports more semantic information from the online encyclopedias. The refined versions of JNet (i.e., JNet1, JNet2 and JNet3) achieve a competitive performance of a coarse version that keeps all the entities without distilling (JNet4) in classification. Moreover, some refined versions of JNet perform outstandingly in clustering. JNet2 performs best in clustering. The best result achieved by JNet2 is 0.822, while the best result achieved by JNet4 is 0.784. The above facts imply that the downsizing of information does not hurt the quality of the ontology and the refined ontologies may even benefit the practical application because they contain less noise information.

6. RELATED WORK

As the most successful online encyclopedia, Wikipedia has attracted much attention from researchers. Abundant semantic relations can be extracted form both the raw text and the structured data of Wikipedia. There are proposed methods based on link analysis between Wikipedia categories [4], using sentence analyzer [6] and taking learning algorithm [23]. Besides, Wikipedia is also used as a rich knowledge bank to enhance other ontology such as WordNet (e.g. [19], [18] and [27]) or enable weakly-supervised learning with automatically generated training examples [2]. Large-scale ontology can be built by restructuring Wikipedia information. Classical works include DBpedia [1] and Yago [20], etc. Yago [20] extracts several dozens of relations from Wikipedia. Yago exploits the redirection system of Wikipedia to extract alternative name of entities and build "*means*" relation. Yago also extracts "*subClassOf*" from the category structure. However, "*subClassOf*" relation is mainly used to connect the leaf categories with WordNet in Yago, while our method focuses on finding the most suitable hypernym-style categories for all entities. The most notable difference be-

[3]http://www.mediawiki.org/wiki/MediaWiki

[4]http://news.sohu.com



(a) Clustering($Purity$) (b) Classification(F^{mi}) (c) Classification(F^{ma})

Figure 2: Clustering and Classification results using different versions of JNet.

tween Yago and our work is that Yago only focuses on Wikipedia while ours integrates various encyclopedia sources and distills the encyclopedic information.

Wikipedia is also used as a bridge cross difference languages [15, 16].As a multi-language semantic knowledge repository with convenient online access, Wikipedia is very suitable for the challenging cross-language tasks. However, there are only a few works focusing on Chinese Wikipedia [24] and automatically building semantic knowledge database from online Chinese encyclopedias [17].

As for entity ranking, previous works mainly focus on better answering a query [22, 11], while our work aims at distilling out more useful and important information from the view of lexical resource construction. For the assessment of article quality of Wikipedia, Hu et al. [8] propose a model making use of interaction between articles and editors acquired from the edit history. Their work takes the words in the articles as the basic units while ours takes the entire article because the encyclopedias we use do not provide such detailed information in the edit history.

7. CONCLUSIONS

In this paper, we propose a method to make better use of the information obtained from different encyclopedias and further align the encyclopedic information with a WordNet-like ontology to construct an enhanced semantic knowledge repository named as JNet. By carefully filtering out the low-quality and less important entities, we can get refined versions of JNet. The experimental results show that the downsizing of the refined version does not hurt the quality of JNet. Furthermore, the refined versions become more practical than the coarse version which retains all the entities from the encyclopedias. To the best of our knowledge, this is the first work about integrating different online encyclopedias to construct an enhanced ontology from the distilled encyclopedic information in Chinese. In general, the method proposed in this paper is not limited to Chinese. However, some technical details may have to be adjusted since Chinese language has its own characteristics.

8. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ive. Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC'07*, pages 722–735, 2007.

[2] L. Bing, W. Lam, and T.-L. Wong. Wikipedia entity expansion and attribute extraction from the web using semi-supervised learning. In *WSDM'13*, pages 567–576, 2013.

[3] C. C. Chang and C.J.Lin. *LIBSVM: a library for support vector machines*. 2001.

[4] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou. Extracting semantic relationships between wikipedia categories. In *SemWiki'06*, 2006.

[5] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *VLDB'04*, pages 576–587, 2004.

[6] A. Herbelot and A. Copestake. Acquiring ontological relationships from wikipedia using rmrs. In *ISWC'06*, 2006.

[7] A. Hotho, S. Staab, and G. Stumme. Wordnet improves text document clustering. In *SIGIR'03*, pages 541–544, 2003.

[8] M. Hu, E. Lim, A. Sun, H. W. Lauw, and B. Vuong. Measuring article quality in wikipedia: Models and evaluation. In *CIKM'07*, pages 243–252, 2007.

[9] S. Jiang, L. Bing, B. Sun, Y. Zhang, and W. Lam. Ontology enhancement and concept granularity learning: Keeping yourself current and adaptive. In *KDD'11*, pages 1244–1252, 2011.

[10] T. Joachims. Training linear svms in linear time. In *KDD'06*, pages 217–226, 2006.

[11] R. Kaptein, P. Serdyukov, A. D. Vries, and J. Kamps. Entity ranking using wikipedia as a pivot. In *CIKM'10*, pages 69–78, 2010.

[12] G. Karypis. CLUTO-a clustering toolkit. Technical report, DTIC Document, 2002.

[13] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *JACM*, 46(5):604–632, 1999.

[14] Y. Liu, S. Yu, and J. Yu. Building a bilingual wordnet-like lexicon: the new approach and algorithms. In *COLING'02*, 2002.

[15] R. Navigli and S. P. Ponzetto. Babelnet: Building a very large multilingual semantic network. In *ACL'10*, pages 216–225, 2010.

[16] T. Nguyen, V. Moreira, H. Nguyen, H. Nguyen, and J. Freire. Multilingual schema matching for wikipedia infoboxes. *PVLDB*, 5(2):133–144, 2011.

[17] J. Nian, S. Jiang, C. Huang, and Y. Zhang. CCE: A chinese concept encyclopedia incorporating the expert-edited chinese concept dictionary with online cyclopedias. *ADMA'11*, pages 201–214, 2011.

[18] S. P. Ponzetto and R. Navigli. Knowledge-rich word sense disambiguation rivaling supervised systems. In *ACL'10*, pages 1522–1531, 2010.

[19] M. Ruiz-Casado, E. Alfonseca, and P. Castells. Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In *AWIC'05*, pages 380–386, 2005.

[20] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A large ontology from wikipedia and wordnet. *Journal of Web Semantics*, 6(3):203–217, 2008.

[21] A. Sumida and K. Torisawa. Hacking wikipedia for hyponymy relation acquisiti. In *IJCNLP'08*, pages 883–888, 2008.

[22] A. M. Vercoustre, J. A. Thom, and J. Pehcevski. Entity ranking in wikipedia. In *SAC'08*, pages 1101–1106, 2008.

[23] G. Wang, Y. Yu, and H. Zhu. Positive-only relation extraction from wikipedia text. In *ISWC/ASWC'07*, pages 580–594, 2007.

[24] R. Wang, Z. Chen, X. Wang, and X. Huang. Analysis on the applications of wikipedia in chinese information processing. In *ICMT'11*, pages 3424–3427, 2011.

[25] G. N. Wilkinson and C. E. Rogers. Symbolic descriptions of factorial models for analysis of variance. *Applied Statistics*, 22:392–399, 1973.

[26] G. N. Wilkinson and C. E. Rogers. Linear models. *Chapter 4 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole*, 1992.

[27] I. Yamada, J. Oh, C. Hashimoto, K. Torisawa, J. Kazama, S. D. Saeger, and T. Kawada. Extending wordnet with hypernyms and siblings acquired from wikipedia. In *IJCNLP'11*, pages 847–882, 2011.

[28] J. Yu, S. Yu, Y. Liu, and H. Zhang. Introduction to Chinese Concept Dictionary. In *ICCC'01*, pages 361–366, 2001.

[29] Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis. *Technical Report TR 01-040, Department of Computer Science, University of Minnesota, Minneapolis, MN*, 2001.