International Journal of Knowledge and Language Processing
Volume 3, Number 3, **2012**

www.ijklp.org
KLP International ©2012 ISSN 2191-2734
pp.1–17

Synchronously Extracting Instances and Attributes for the Concepts from the Web

Sui Zhifang[1], Kang Wei[2] and Tian Ye[1]
[1] Peking University
No.5, Yiheyuan Road, Haidian District, Beijing, China
szf@blcu.edu.cn

[2] Shanghai Research Institute of China Telecom

Received June 2012; revised September 2012

ABSTRACT. *As an important task in ontology learning, instance extraction and attribute extraction of the concepts have attracted more research attention. There are close relations between the instances and the attributes of a concept. Detecting an instance will be beneficial for recognizing attributes, and vice versa. This paper puts forward a weakly-supervised method which can synchronously extract instances and attributes for a concept based on web information. Firstly, we automatically generate and evaluate the contextual patterns in which instances and attributes co-occur on the Web. Secondly, we extract candidate instances and attributes using the patterns extracted above, and evaluate them with two methods. (1)Based on the associations between instances and attributes, we use determinate instances (i.e. the seed instances) to evaluate the accuracy of the candidate attributes; and use the determinate attributes (i.e. the seed attributes) to evaluate the accuracy of the candidate instances. (2) We use the contextual distribution similarity to evaluate the accuracy of instance extraction and attribute extraction. Experiment results show that we could get a comparable performance with the traditional methods, but from the corpus with much less size.*

Keywords: Instance Extraction; Attribute Extraction; Ontology Learning; Weakly-supervised Method

1. **Motivation.** Ontology supports the sharing and reuse of knowledge across different applications. However, manual ontology construction costs a lot of manpower, material and financial resources, which hampers the application of ontology to some extent. Therefore, automatically constructing ontology has become a research hotspot currently.

As an important task in ontology learning, instance extraction and attribute extraction of concepts have attracted more research attentions. The task of instance extraction is:for each given concept c, acquire an instance set $I=\{i\}$, in which each i is an instance of c. For example, *Influenza* and *Hypertension* can be regarded as the instances of the concept *DISEASE*. On the other hand, the task of attribute extraction is:for each given concept c, acquire an attribute set $A=\{a\}$, in which each a is an attribute of c. One concept differs from the other in that they have different attributes. For example, the concept *DISEASE*

has the attributes as *SYMPOTOM* and *TREATMENT* etc., while the concept *MEDICINE* takes the attributes of *SIDE-EFFECT* and *EFFICACY*, etc.

r*The methods for instance extraction.* [1] extracted *INSTANCE-OF* relationships from the texts using the syntactic patterns such as "such NP_o as NP_i, . . . ,NP_{n-1} (or|and) other NP_n". They obtained good accuracy, however the extracting-from-text method caused the data sparseness problem. Other researches extracted instances from the unstructured web corpus using unsupervised or weakly supervised methods [2-6]. However, with little hand-labeled training examples, it's hard to improve both recall and precision.

The methods for attribute extraction.[7] extracted candidate attributes of a concept from the Web and used two supervised classifiers to determine whether a candidate is a correct attribute. To construct the classifiers, they used the features of morphological information, an attribute model, a question model and an attributive-usage model. [8] used an unsupervised method to extract attribute-value pairs from semi-structured HTML web pages. [9] used a weakly-supervised method to extract attributes of a concept from structured web pages. In recent years, with the development of Wikipedia, Wikipedia-based attribute extraction method has attracted more and more attention [10].

[11] first defines a closed vocabulary of potential class instances as the set of most frequently-submitted Web Search queries, then acquires class labels for potential class instances via hand-written extraction patterns as the form of <C[such as|including]I>, and organizes potential class instances into sets of distributional similar phrases. Then he uses TF-IDF (term frequency–inverse document frequency) to rank candidate <class label, instance> pairs. In the attribute extraction part, he uses similarity-based ranking method which is introduced in his previous work [12]. This work also compared different similarity functions such as Cosine, Jacard, Jensen-Shannon, and Skew-Divergence. Most of the above researches pay attention to separate instance extraction or attribute extraction. As an exception, [11] used a weakly-supervised method to synchronously extract instances and attributes for open domain concepts from web pages and search engine query logs. However, they used 50,000,000 query logs and 100,000,000 web pages. It is difficult to obtain so large scale of query logs and the web pages they used are tremendous. In this paper, we make good use of the close relationship between the instances and the attributes of a concept to extract both of them. Using only the titles and snippets of the web pages returned by Google as the corpus, we could reach a comparable performance with [11].

Since attributes are the intrinsic characteristics of a concept, they are often used to describe an instance of a concept. For example, if we expect to find out *influenza*, an instance of the concept *DISEASE*, we usually concern its symptoms, its treatments, etc. Therefore, *influenza* often co-occurs with *symptom*or*treatment* in its context. This phenomenon is prominent in the Web. There are tremendous web pages which mention a concept instance and its attributes simultaneously. Therefore, if we search the Web using an instance of a concept as the query, we may probably find that the instance and the attributes of the concept co-occur in the returned web pages; moreover, the co-occurrence patterns tend to be stable. For example, "肾结石(kidney stone) 的(of) 治疗方法(treatment)", "心绞痛(angina) 的(of) 临床表现(clinical manifestation) 是(is)", "感冒

(influenza) 的(of) 症状(symptom)有(includes)". Among them, "肾结石(kidney stone)", "心绞痛(angina)"and "感冒(influenza)" are the instances of the concept "疾病/DISEASE"; while"治疗方法/treatment", "症状/symptom" and "临床表现/clinical manifestation" are the attributes of the concept 疾病/DISEASE. Therefore, there is close relationship between the instances of a concept and its attributes. Detecting instances of a concept will benefit the recognition of its attributes, and vice versa.

Based on this motivation, this paper puts forward a weakly-supervised method which can synchronously extract instances and attributes for a concept based on the web information.

2. **Methodology.** The task of synchronously extracting instances and attributes for the concepts of ontology is defined as follows: For a concept *c*, given seed instances and seed attributes, we expect to extract from the Web the instance set *I* and the attribute set *A* of *c*. Our system consists of five components: contextual pattern extraction and evaluation, instance extraction, instance evaluation, attribute extraction and attribute evaluation. The architecture is shown as Figure 1.

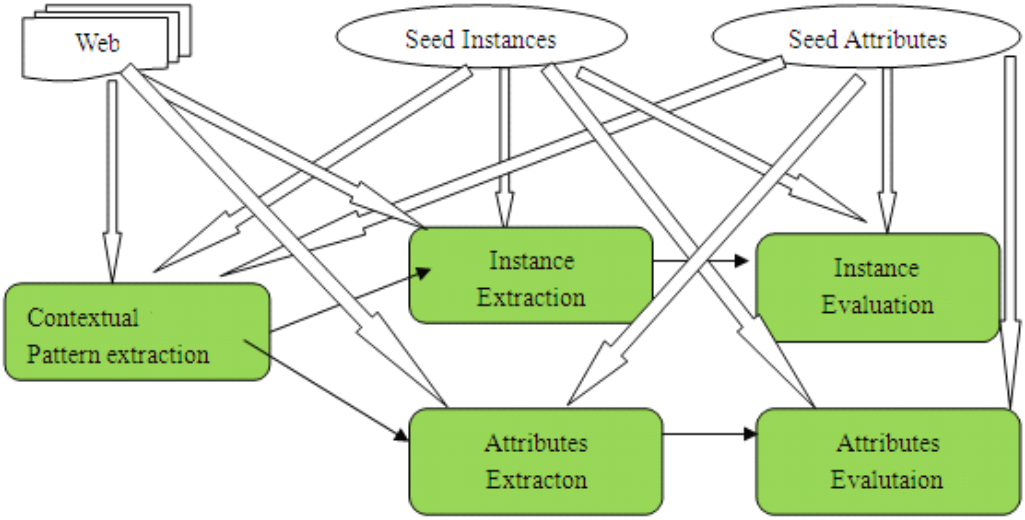

FIGURE 1. Architecture of our method

(1) **Contextual pattern extraction and evaluation.** We use the seed instances and the seed attributes to construct queries to search the Web. Then, from the returned web pages of Google search, we extract the contextual patterns in which the instances and the attributes co-occur. Finally we evaluate these patterns using the seed instances and the seed attributes.

(2) **Instance extraction.** We use the seed attributes and the contextual patterns extracted in the first module to construct search queries, and use them to search the Web and extract the candidate instances from the returned web pages.

(3) **Instance evaluation.** We evaluate the candidate instances with two methods.

(a) Calculating the association between the candidate instances and the seed attributes.

(b) Calculating the contextual distributional similarity between the candidate instances and the seed instances.

After evaluation, we can expand the seed instance set using the credible candidate instances.

(4) **Attributes extraction.** We construct search queries using the contextual patterns obtained in the first module and the expanded seed instance set and extract the candidate attributes from the returned web pages of Google search.

(5) **Attributes evaluation.** We evaluate the candidate attributes with two methods.

(a) Calculating the association between the candidate attributes and the seed instances.

(b) Calculating the contextual distributional similarity between the candidate attributes and the seed attributes.

After evaluation, we can expand the seed attribute set using the credible candidate attributes.

We will describe the above modules in detail in the following sections.

3. Extraction of the Contextual Patterns

3.1 Extraction of Contextual Patterns. The relationship between an instance i and an attribute a is practically a kind of relation of "a is an attribute of i". For example, "the symptom of influenza includes" and "the treatment of hypertension requires" respectively embody the relationship of "symptom is an attribute of influenza" and "hypertension has an attribute of treatment". Therefore, we try to extract the contextual patterns like "iH1aH2", where i is an instance, a is an attribute, H1 and H2 are contextual snippets whose frequencies are higher than a threshold F and whose lengths are lower than a threshold L. The threshold of F is set to 100 and L is set to 5 which can filter out noisy contextual patterns according to our experiments.

For example, for the concept of *DISEASE*, given the seed instance "感冒/influenza" and the seed attribute "症状/symptom", we can construct the query of "感冒*症状*", then we can search the Web using the query of "感冒 * 症状*".

Using Google search, we can obtain at most the top 1000 web pages for each search query. We only use the top 500 searching result for each query. Instead of downloading these web pages, we only use the titles and the snippets of these web pages to compose the corpus for contextual pattern extraction, denoted as corpus *P*. Then, we use the seed instances and the seed attributes to extract the contextual patterns like iH_1aH_2, the obtained pattern set is denoted as *Y*.

3.2 Evaluation of Contextual Patterns. In the above module, we have extracted the contextual pattern set *Y* from the corpus *P* and obtained the frequencies in which each pattern γ occurs in the corpus *P*. However, we cannot evaluate the pattern γ simply by its frequency, since the probability that γ occurs in the corpus *P* is insufficient in determining the effectiveness of the pattern. Instead, it will be more accurate if we use the probability in which γ occurs in the Web, which is calculated as follows.

$$score(\gamma) = \sum_{<i,a>\in(I\times A)} p(\gamma \mid <i,a>)$$
$$= \sum_{<i,a>\in(I\times A)} \frac{p(<i,a>,\gamma)}{p(<i,a>)}$$
$$= \sum_{<i,a>\in(I\times A)} \frac{Hits(<i,a>,\gamma)}{Hits(<i,a>)} \quad (1)$$

In the equation, ($<i, a>$, γ) denotes the operation of replacing i and a in the pattern iH_1aH_2 with the specific instance i and a, Hits(q) denotes the number of search result if we use q as the Google query to search the Web. We select the top 10 patterns, which could generate balanced precision and recall, and obtain the final pattern set Y', and normalize the weight of γ in Y' as

$$score'(\gamma) = \frac{score(\gamma)}{\sum_{\chi\in Y'} score(\chi)} \quad (2)$$

Table 1 lists the top 5 contextual patterns and their weights. In the following we will use the pattern set Y' to extract instances and attributes from the Web.

TABLE 1. Top-5 contextual patterns

	Contextual patterns	Examples	Weight
1	*i*+的+*a*+方法 (的/'s, 方法/method)	肾结石的治疗方法(the *treatment* of *kidney stone*) 肺结核的诊断方法(the *diagnostic method*of *phthisis*) 高血压的保健方法(the *health-care measure* of *phthisis*)	0.3289
2	*i*+的+*a*+及 (的/'s, 及/and)	高血压的预防及(the *precaution*of*phthisis* and) 糖尿病的饮食及(the *diet* of*diabetes* and) 鼻炎的症状及 (the *symptoms* of*rhinitis*and)	0.0998
3	*i*+的+*a*+有 (的/'s, 有/include)	感冒的症状有(the *symptoms* of *influenza* include) 肾结石的病因有(the *pathogenesis* of *kidney stone* include) 抑郁症的表征有(the *characterizations* of *depression* include)	0.0781
4	*i*+的+*a*+是 (的/'s, 是/be)	鼻炎的症状是(the *symptoms* of *rhinitis* are) 心绞痛的临床表现是(the *clinical manifestations* of *angina pectoris* are) 糖尿病的前兆是(the *precursors* of *diabetes* are)	0.0649
5	*i*+的+*a*+和 (的/'s, 和/and)	感冒的预防和(the *precaution* of *influenza* and) 心绞痛的诊断和(the *diagnose* of *angina* and) 颈椎病的用药和(the *medication* of *cervical spondylosis*and)	0.0624

4. Extraction of Instances. In this section, we will use the contextual patterns set Y' obtained in section 3 and the initial seed attributes to construct queries, and extract candidate instances from the returned web pages of Google. Furthermore, we propose two approaches to evaluate the credits of the extracted instances.

4.1 Extraction of Candidate Instances. The process of extraction of candidate instances

is also based on Web searching. Queries are built with the contextual patterns and individual seed attributes. In the contextual pattern,we replace *a* with a specific attribute and construct the query *H_1aH_2. We can obtain the web snippets which meet the above pattern through searching the query in Google. We use the seed attribute set Δ and the contextual pattern set Y to construct searching query set Q.

For example, for the pattern "*i* 的/'s *a* 和/and" and the seed attribute "症状/symptom" we construct query "* 的/'s 症状/ symptom 和/and", and obtain the titles and snippets of the web pages returned by Google, which compose corpus I.

We use the following strategy to extract instances from corpus I based on the pattern set Y. First, we extract sentences from corpus I. Then we build search queries with the contextual patterns and the seed attribute. Queries in a pattern of "H_0-*-H_1-attribute-H_2", where H_0 stands for left-context, do not work well since that the instances in the Web pages tend to occur at the beginning or in the front positions of the sentence of the result web pages. Therefore, we only select the sentences whose beginning parts agree with the pattern "*H_1aH_2", and extract the chunks matching "*". These chunks compose the set S. After that, we trim the Chinese character sequences (chunks) in S. Specifically we eliminate the noise prefixes and suffixes depending on a prefix stop list and a suffix stop list, and we only retain the word sequences with the lengths between 2 to 10 characters. For example, we get a sentence "现在(now)中国(China)的('s)首都(capital)是(is)哪里(where)?"After matching the pattern, we extract "现在中国/China now" matching "*" and "现在/now" will be dropped as stop word, so we get "中国/China" as a candidate instance. Through the above filtering, finally we obtain the candidate instance set.

4.2 Evaluation of Candidate Instances. It is inevitable that the candidate instances extracted using the contextual patterns include some noises. Therefore, we need to evaluate the confidence of each candidate. We propose two approaches to conduct the evaluation.

4.2.1 Evaluation Approach Based on PMI-IR. An authentic instance should be highly associated with the seed attributes. Therefore, we use the mutual information between the candidate instance and the seed attributes to measure the confidence of the candidate instance. Because the confidence of the seed attributes is known to be 1, we can use the confidence of the seed attributes to compute the confidence of the candidate instance. In the paper, we use PMI-IR (Pointwise Mutual Information and Information Retrieval) to compute *Weight*(*i*) of the candidate instance:

$$Weight(i) = \frac{\sum_{a \in A} \left(\frac{pmi(i,a)}{\max_{pmi}} * Weight(a) \right)}{|A|} \qquad (3)$$

$$pmi(i,a) = \log \frac{Hits(i,a) * N}{Hits(i) * Hits(a)} \quad (4)$$

In the equation, *Weight*(*i*) denotes the confidence of the candidate instance *i*; *Weight*(*a*) denotes the confidence of the seed attribute *a*, and $\max_{pmi}$ is the maximum pointwise mutual information between candidate instance and all seed attributes. *Hits*(*q*) denotes the number of search result if we use *q* as the query to search the Web. (*i*, *a*) denotes that both *i* and *a* are used as the words of query, *N* denotes the number of the pages in the Web.

4.2.2 Evaluation Approach Based on Similarity. Besides that the correct candidate instances are more closely related to the attributes of the concept, the contextual patterns of the correct candidate instances are also more similar with that of the seed instances. Therefore, we can measure the similarity between the candidate instances and the seed instances in order to evaluate the confidence of the candidate instances. The approach consists of three steps.

(1) Based on the contextual patterns extracted by *Y'* in Section 3, we construct a feature vector for each seed instance. Specifically, for each contextual pattern γ ="$\iota H_1 a H_2$"in *Y'*, we replace *i* and *a* with a specific seed instance and a seed attribute, use it as the query to search the Web and obtain *Hits*(γ). Then, the weight for each feature in the feature vector is calculated as follows:

$$p(\alpha \mid \gamma) = \frac{p(\alpha,\gamma)}{p(\gamma)} = \frac{Hits(\alpha,\gamma)}{score'(\gamma) * N} \quad (5)$$

In this equation*Hits*(α,γ) is the count of hits usingα and γ as the query words. *score'*(γ) is the weight of the pattern γcalculated in section 3.2, and *N* denotes the number of pages on the Web.

After we generate the feature vector for each seed instance, the sum of the feature vectors are normalized by the number of feature vectors and generate a reference feature vector *vs*.

(2) With a similar approach, we construct a feature vector for each of the candidate instances i_{cand}. In this way, we can obtain the feature vector *vc* for each candidate instance.

(3) We compute the similarity between the feature vector of each candidate instance *vc* and the reference feature vector *vs*, and rank the candidate instances by their similarity. In our work Jensen-Shannon divergence [13] is used to compute the similarity, *Sim*(*vc*,*vs*). Jensen-Shannon divergence is a metric which measures the distance between two distributions.

$$JS(q,r) = \frac{1}{2}\left[D\left(q \| avg_{q,r}\right) + D\left(r \| avg_{q,r}\right)\right] \quad (6)$$

$$D\left(\left(p_1(V)\right)\middle\|\left(p_2(V)\right)\right)=\sum_{v} p_1(v)\log\frac{p_1(v)}{p_2(v)} \qquad (7)$$

After these steps, the candidate instances are ranked by their similarity with the seed instances.

5. Extraction of Attributes. Similar to instance extraction, attribute extraction is also decomposed into candidate attribute extraction and candidate attribute evaluation. For the evaluation of candidate attributes, we use the approaches similar to candidate instance evaluation, i.e. the approach based on the association between the candidate attribute and the seed instances and the approach based on the contextual distributional similarity between the candidate attribute and the seed attributes.

5.1 Extraction of Candidate Attributes. In Section 4.1, we use the seed attributes and the contextual pattern set V to construct query to extract the candidate instances. For attribute extraction, however, it is infeasible to use seed instance and the contextual pattern set alone, since the number of attributes is far less than the number of instances. Only using the seed instances to search the Web may result in even fewer attributes. Therefore, we add the highly ranked candidate instances from section 4 to the seed instance set, and combine the expanded seed instances with V to construct a query set Q to extract candidate attributes.

By searching the Web with query Q, we can get the titles and the snippets of the returned web pages, denoted as corpus A. For extracting the attributes from the sentences of corpus A, we select the sentences which have the patterns of H_1*H_2 and extract the text spans which match "*", denoted as S. Only word sequences with lengths between 2 and 10 characters are retained. Through the above filtering, finally we can obtain the candidate attributes set.

5.2 Evaluation of Candidate Attributes.

5.2.1 Evaluation Approach Based on PMI-IR. The evaluation approach based on the PMI-IR is similar to the evaluation of candidate instances, except that in the evaluation of candidate instances we add the instances whose confidence is higher than a threshold into the seed instance set. The following equation defines the evaluation of the confidence of the candidate attribute a.

$$Weight(a)=\frac{\sum_{i\in I}\left(\frac{pmi(a,i)}{\max_{pmi}}*Weight(i)\right)}{|I|} \qquad (8)$$

In this equation $pmi(a,i)$ is identical to $pmi(a,i)$ in equation(4).

5.2.2 Evaluation Approach Based on Similarity. The confidence of the candidate attributes are also evaluated by their similarity with the seed attributes. The similarity is calculated through the following three steps.

(1) We construct the reference feature vector using the seed attributes. The construction method is similar to that described in section 4.2.2, except that here we replace the seed instances with the seed attributes.

(2) We construct the feature vectors for the candidate attributes. The construction method is equal to that in section 4.2.2.

(3) We compute the similarity between the feature vectors of the candidate attributes and the reference feature vector, and rank these candidates by their similarity.

After these steps, the candidate attributes are ranked by their similarity with the seed attributes.

6. Experiments and Analysis

6.1 Data Sets. We use Google as the search engine for obtaining the Web corpus. After constructing search queries, we used the titles and the snippets of the web pages returned by Google as the corpus for extracting the contextual patterns, instances and attributes. The difference between our method and the method of [11] lies in that they use large scale query logs and web pages as the corpus, while we use the top 500 web pages returned by Google for each query, making use of only titles and snippets rather than downloading the full pages. Therefore, the size of the corpus we used is far less than that of [11].

The experiments are conducted separately in Chinese corpus and English corpus. As for Chinese, we conduct experiment using 20 concepts, such as “Mountains”, ”Awards”, “NBA Teams”, “Painters”, “Countries”, “Car Models”, “Religions” and so on. There are thirteen testing classes which are same with [11]. For English, we conduct experiment using the concept of “Company” and “Country”, which is the same with the experiments in [3, 11, 14].The English gold standard comes from the en.wikipedia.org and rest of the Chinese gold standard comes from the zh.wikipedia.org and baike.baidu.com.

We use precision and coverage to evaluate the results of instance extraction and the results of attribute extraction. For the evaluation of precision, we used human determination method, i.e. for each extracted instance we manually determined whether it belongs to the given concept or not. Since the exact recall of the extracted instances and attributes is impossible to be given, in this paper we use the metric of coverage instead of recall. Here we define the coverage as the ratio of the interaction between the extracted instances/attributes and the instances/attributes in the gold standard.

6.2 Results of Instance Extraction. We use the contextual patterns, the seed instances and the seed attributes to construct the search queries, and extract candidate instances from the corpus composed by the search results of Google. Then, we evaluate the confidence of the candidate instances based on (1) the PMI-IR between the candidate instance and the seed attributes, and (2) the similarity between the contextual pattern similarity of the candidate instance and the seed instances.

6.2.1 Coverage of Instance Extraction. The instance extraction results for the concepts are shown in Table 2.

TABLE 2. Comparison between instance sets of gold-standard classes (G) and instance sets of automatically-extracted (E).

	Concept	Size of instance sets		
		G	E	G∩E/G(%)
Chinese	NBA 球队/NBA team	30	74	100.00
	车型/Car model	395	801	66.84
	国家/Countrie	196	754	84.18
	河流/River	756	405	19.74
	花/Flower	547	890	16.09
	画家/Painter	760	706	10.39
	疾病/Disease	900	2202	39.20
	奖项/Award	630	633	49.68
	节日/Holiday	630	1387	15.55
	美国州/State(U.S.A)	50	142	92.00
	山峰/Mountain	706	962	15.01
	诗人/Poet	1326	936	11.71
	史书/Historical Record	191	625	23.56
	药物/Drug	1576	1128	27.00
	运动员/Athlete	759	539	5.79
	哲学家/Philosopher	535	483	12.71
	中国朝代/Dynasty in Chinese history	79	657	54.43
	中国省份/Province of China	23	45	100.00
	宗教/Religion	153	228	20.91
	作曲家/Composer	554	451	11.37
English	Country	196	225	60.00
	Company	500	4353	42.00
	Average coverage			39.92

Each gold-standard instance set comes from the web encyclopedia (zh.wikipedia.org or baike.baidu.com). Ratios (G∩E/G) are shown as percentages.

6.2.2 Precision of Instance Extraction. In the experiments, we evaluate the precision of instance extraction with the human-constructed gold standards for 4 concepts of "疾病/Disease", "药物/Drug", "公司/Company" and "国家/Country", among which"疾病/Disease", "药物/Drug", "公司/Company" are open sets, so the instances are not fully covered by gold standard knowledge base; the instances of concept "国家/Country" is closed, covered by the knowledge base. The precision is shown in Table3, in which the last column refers to the precision of the top 100 candidate instances.

TABLE 3. The precision of instance extraction

Concept	Seed instances	Seed attributes	Top 5 candidate instances	Precision
Disease (Chinese)	鼻炎/ rhinitis 肾结石/ kidney stone 颈椎病/ cervical spondylosis 高血压/ hypertension 感冒/ influenza	病因/ cause of disease 治疗/ treatment 症状/ symptom	牙结石/ dental calculus 放射性肠炎/ radiation enteritis 白塞病/ Behcet'sdisesase 雷诺综合征/ Raynaud syndrome 内分泌失调性不孕/ endocrine infertility	92%
Drugs (Chinese)	阿替洛尔/ atenolol 硝化甘油/ nitroglycerin 奎尼丁/ quinidine 洋地黄/ digitalis 地高辛/ digoxin	禁用慎用/ disable and be used with caution 用法用量/ indication et dosage	埃索美拉唑镁肠溶片/ esomeprazole magnesium enteric-coated tablets 精氨酸阿司匹林/ arginine aspirin tablets 匹多莫德分散片/ pidotimod tablet 替卡西林/ ticarcillin 哌拉西林/ piperacillin	67%
Company (English)	IBM Google GE	headquarter revenue founder	Citigroup Nike IBM Coca cola Wal-mart	88%
Country (English)	France China England Canada Japan	Capital Population National symbols	France China India Canada Thailand	91%

In the experiments of Chinese, the precision of the top 100 candidate instances of the concept "疾病/Disease" is above 92%, covering 39.2% of the 900 common diseases in the gold standard. The same result of the concept "药物/Drugs" is above 67%, covering 5.83% of the 1,500 common drugs in the gold standard. The lower performance than that of "疾病/Disease" may due to that the attributes of the concept "疾病/Disease" are more specific in comparison with those of "药物/Drugs". For example, "病因/Cause of disease", "症状/Symptom" and "治疗/Treatment" are all the attributes specific to the concept of "疾病/Disease", while the attribute "副作用/Side effect" is not only the attribute of concept "药物/Drugs" but also that of concept "手术/Operation", etc.

In the experiments in English, the precision for the top 100 candidate instances of the concept "Company" is 88%, covering 42% of the 500 companies in the gold standard. For concept "Country", the precision achieves 91%, the coverage achieving 60%.

The performance of instance extraction in the two English concepts is superior to that of the two concepts in Chinese. The reason probably lies in that (1) the instance set of

TABLE 4. The precision of instance extraction for concept“疾病/Disease”

N	Precision of instance, top N
100	92%
200	94%
500	94%
1,000	93%
2,000	85%

concept “Country” is closed, while the instance sets for other three concepts is open; (2) no word segmentation problem exists in English. So the candidate instances extracted are relatively complete linguistic units.

We conducted more detailed experiments for the concept of “疾病/Disease”. Table 4 shows the precision of instance extraction for concept “疾病/Disease”. Through human determination, we find that the precision of top500 and top1,000 candidate instances achieves 94% and 93%, respectively.

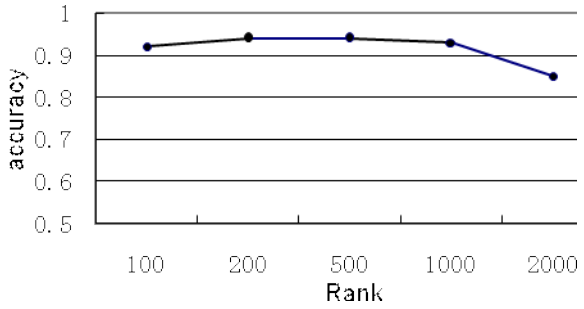

FIGURE 2. Precision of instance extraction for concept “疾病/Disease”

Figure 2 shows that the precision of the top 1,000 instance candidates after ranking achieves 92%, and that of top 2,000 still remains at 85%. The reason that the precision is decreased is that after ranking, the instance candidates with higher confidences are mostly ranked in the front positions, while those with lower confidences are ranked in the back low.

6.2.3 Comparison with Other Methods. Table 5 shows the comparison with other methods. [3] used human selected pattern to extract instances for concept “Company” and “Country” from the web page corpus with the size of 60 million words, and they evaluate the precision for instance extraction using a sampling and human determination method. They extracted 1,116 instances for concept “公司/Company” , with the precision of 90%. [14] used an unsupervised method to extract the instances for the concept of “Country”, “constellation” and “fish species” from the unstructured web corpus. The average precision is 81%. [11] extracted 40 classes of instances, and the average accuracy is 26.89%. Accuracy Ratios used in [11] is identical to coverage ratio in our result. In comparison with the above three methods, our method extract instances with a comparable precision and coverage but much less corpus.

TABLE 5. The comparison with other methods

The methods	Number of concepts	Precision	Coverage
[3]	2 classes	90%	N/A
[14]	3 classes	81%	83%
[11]	40 classes	80%	26.89%
Ours	22 classes	84%	39.92%

6.2.4 Influence of Different Seed Attributes over the Performance of Instance Extraction. Table 6 shows that the selection of different seed attributes will have an impaction the precision and coverage of instance extraction. In the first experiment for concept "药物/Drugs", the seed attribute "副作用/Side effect" and "适应症/Indication" are not concept-specific. They are also the attributes of concepts e.g."手术/Operation" and "治疗方法/Therapeutic method". Therefore, the candidate instances like "颞部填充术/Temporal filling", "眉部整形/Eyebrow shaping", "脂肪抽吸术/Liposuction" are extracted. In the second experiment, we used the seed attributes like "禁用慎用/ Disable and be used with caution" and "用法用量/Indication et dosage", which are more specific to concept "药物/Drugs". The result shows that the precision for instance extraction has been obviously improved in comparison with the first experiment.

On the other hand, because the contextual patterns for the seed attributes "禁用慎用/ Disable and be used with caution" and "用法用量/Indication et dosage" are relatively less, the coverage of instance extraction is decreased. Therefore, selecting more specific seed attributes is important for improving the precision for instance extraction.

TABLE 6. The influence of different seed attributes over the precision of instance extraction

Seed attributes.	Top 5 instances extracted	Precision
副作用/Side effect 适应症/Indication 药理作用/Pharmacological actions	灯盏花素/Breviscapinun 颞部填充术/temporal filling 降血糖药/Metformin 枸杞/medlar 复方丹参片/compound salvia tablet	48%
禁用慎用/ Disable and be used with caution 用法用量/ Indication et dosage	埃索美拉唑镁肠溶片/ Esomeprazole magnesium enteric-coated tablets 精氨酸阿司匹林/Arginine aspirin Tablets 匹多莫德片/Pidotimod tablet 替卡西林/Ticarcillin 哌拉西林/Piperacillin	67%

6.2.5 Comparison with Gold Standard Knowledge Base. Experiment results suggest that there is a common phenomenon for the testing concepts that the precision is relatively high while the coverage is relatively low. Moreover, Table 7 shows that many correct instance candidates are not included in the gold standard knowledge base, which suggests a method to expand the gold standard knowledge base to some extent. The negative part of the results is related to the seed attributes that we choose. Though the incorrect instances are extracted, they positioned the bottom of the result list of candidate instances after ranking.

TABLE 7. Comparison with gold standard knowledge base

Concept	Correct candidates covered by gold standard	Correct candidates not covered by gold standard	Incorrect candidates not covered by gold standard
疾病/ Disease (Chinese)	58 instances	1810 instances 胃窦炎/Antritis 溶血症/haemolyticus 坐骨神经痛/sciatica	410 instances 骨刺病人/ Spur patient 血虚患者/ Patients with blood deficiency
药物/ Drugs (Chinese)	42 instances	310 instances 复方丹参片/ compound salvia tablet 清肺散结丸/ Qingfeisanpill	600 instances 颞部填充术/ temporal filling 脂肪抽吸术/ Liposuction
	9 instances	91 instances 替卡西林/Ticarcillin 奥卡西平 /Oxcarbazepine 左乙拉西坦/ Levetiracetamtablets	100 instances 非处方药 /Over-the-counter medicine 适应症 /Indication 儿童用药 /Medication in infants and children
Company (English)	210 instances	0 instances	39 instances Communist Party Fudan
Country (English)	106 instances	0 instances	11 instances Wisconsin Rome New Jersey

6.3 Results of Attribute Extraction

6.3.1 Attribute Extraction Results for the 4 Concepts. The attribute extraction results for 4 concepts are shown in Table 8.

TABLE 8. The attribute extraction results for the 4 concepts

Concepts	Seed attributes	Top 5 attributes extracted	Precision		Coverage
疾病/ Disease (Chinese)	病因/ cause of disease 治疗/ treatment 症状/ symptom	病因/ cause of disease 体征/ physical sign 发病部位/ location of the diseases 疗效/ curative effect 用药/ Medications	Top5	80%	70.60%
			Top10	60%	
			Top15	67%	
			Top20	70%	
药物/ Drugs (Chinese)	副作用/ side effect 适应症/ indication	药理作用/ pharmacological actions 不良反应/	Top5	80%	75%
			Top10	70%	
			Top15	73%	
			Top20	70%	

	药理作用/ pharmacological actions	adverse reaction 食品/ food 副作用/side effect 制备/preparation			
Company (English)	headquarter revenue founder	CEO application Chairman creation division	Top 5	80%	77%
			Top 10	80%	
			Top 15	80%	
			Top 20	85%	
Nation (English)	Capital Population religion	economy center Government population Society	Top5	100%	91%
			Top10	80%	
			Top15	73%	
			Top20	80%	

Figure 3 shows the precision of attribute extraction for the concepts of "疾病/Disease", "药物/Drugs", "公司/Company" and "国家/Nation".

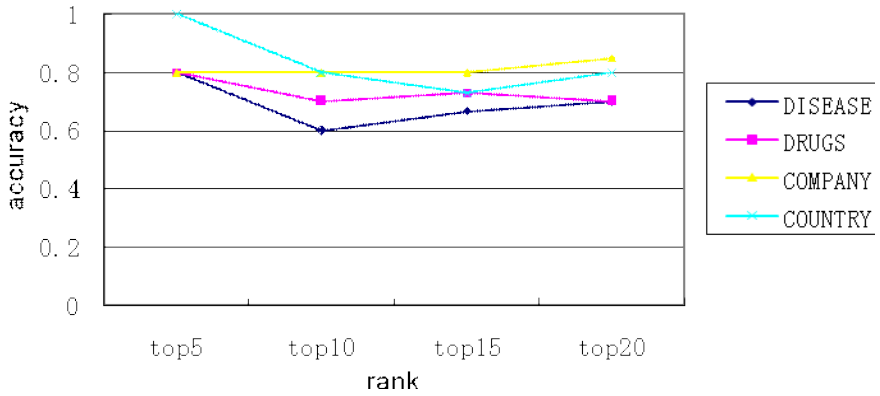

FIGURE 3. Precision of attribute extraction for the 4 concepts

In the experiments in Chinese, the top-5 precision for the attributes extraction of both the concept "疾病/Disease" and "药物/Drugs" is over 80%, and top-20 precision achieves 70%. The coverage of the concept "疾病/Disease" and "药物/Drugs" achieves 70.6% and 75%, respectively.

In the experiments in English, the top-5 precision for "Company" is over 80%; and the top-5 and top-20 precision for "Nation" achieves 100% and 80%, respectively. The coverage of attribute extraction for the concepts of "Company" and "Nation" achieves 77% and 91%, respectively.

Similar to instance extraction, the performance of attribute extraction in English is superior to that in Chinese. The reason probably lies in that no word segmentation problem exists in English. So the candidate attributes extracted are relatively complete linguistic units.

6.3.2 Comparison with other Methods. [11] used an unsupervised method to extract the attributes of the concepts of open domains from the Web corpus and search engine logs, the precision for the top-20 candidate attribute achieves 67%. In comparison with the 50

million query logs and 100 million web pages they used, the resources we used are much less. However, our method can achieve comparable precision with their method.

6.3.3 Comparison with Gold Standards. Table 9 shows the comparison with gold standard knowledge base. Many correct attribute candidates are not included in the gold standard knowledge base, which suggests a method to expand the gold standard knowledge base.

TABLE 9. Comparison with gold standard knowledge base

Concept	Correct candidates covered by gold standard	No. of the correct extracted attributes which are not coincided with gold standard, and examples	No. of the incorrect extracted attributes which are not coincided with gold standard, and examples
疾病/ Disease (Chinese)	12 instances	5 instances 防治/ prevention and cure 诊断标准/ standard of diagnosis 影像检查/ imaging	40 instances 根治/ radical treatment 规范化/ normalization
药物/ Drugs (Chinese)	9 instances	13 instances 不良反应/ untoward effect 制备/ preparation 作用机制/ mechanism of action	33 instances 食品/food 化疗方案/ GEMCAP 分子量/ molecular weight
Company (English)	10 instances	45 instances Co-CEO Chief Information Officer	25 instances Application Result List
Country (English)	11 instances	15 instances laws border areas	25 instances role importance issue

7. Conclusions. This paper proposes a method of synchronously extracting instances and attributes for the concepts based on the Web. The evaluation shows that appropriate selection of seed attributes can achieve notable average precision and coverage of instance extraction on testing datasets. However, what kind of attributes is appropriate for the extraction of a certain concept? By experiment we suggest that the appropriate attributes are the most specific ones which could distinguish the current concept from others. For example, the concept of “Poet”, “Athlete”, “Painter” and “Philosopher” are all sub-concepts of “People”, we use the seed attributes set <诗作/poem, 诗意/poetry> to distinguish “Poet” from other sub-concepts of “People”. So a concept hierarchy should be helpful for selecting the appropriate seed attributes for the instance extraction. As for future works, we will try to conduct some experiments on this idea.

REFERENCES

[1] HEARST, M. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, 1992, 539-545.

[2] ETZIONI, O., CAFARELLA, M., DOWNEY, D., POPESCU, A., SHAKED, T., SODERLAND, S., WELD, D.S., AND YATES, A. 2005, Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*165, 91-134.

[3] CAFERELLA, M., DOWNEY, D., SODERLAND, AND S., ETZIONI, O. 2005. KnowItNow.Fast, Scalable Information Extraction from the Web. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing(HLT/EMNLP)*, 2005.Vancouver, 563-570.

[4] KELLER, F., LAPATA, M., AND OURIOUPINA, O. 2002.Using the Web to Overcome Data Sparseness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002, Philadelphia, 230-237.

[5] BLOHM, S., AND CIMIANO, P. 2007. Using the Web to Reduce Data Sparseness in Pattern-Based Information Extraction. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, 2007, Warsaw, Poland,18-29

[6] TURNEY, P. 2001. Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning (ECML-2001)*, 2001, Freiburg, Germany, 491–502.

[7] POESIO, M., AND ALMUHAREB, A. 2005.Identifying Concept Attributes Using a Classifier. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, Ann Arbor, 2005, 18-27.

[8] YOSHINAGA, N., AND TORISAWA, K. 2007.Open-Domain Attribute-Value Acquisition from Semi-Structured Texts. In *Proceedings of the OntoLex 2007 – From Text to Knowledge: The Lexicon/Ontology Interface*, Busan, South-Korea, 2007.

[9] RAVI, S., AND PASCA, M. 2008.Using Structured Text for Large-Scale Attribute Extraction. In *Proceedings of the 17th International Conference on Information and Knowledge Management(CIKM-08)*, 2008, Napa Valley, California, USA,1183-1192.

[10] CUI, G., LU, Q., LI, W., AND CHEN, Y. 2009.Automatic Acquisition of Attributes for Ontology Construction. In: *ICCPOL2009*, Vol.5459, Springer(2009), 248-259.

[11] PASCA, M., AND DURME, B.V. 2008.Weakly-Supervised Acquisition of Open-Domain Classes and Class Attributes from Web Documents and Query Logs. In *Proceedings of the ACL-08: HLT*, 2008, Columbus, Ohio, USA, 19-27.

[12] PASCA, M. 2007. Organizing and Searching the World Wide Web of Facts- Step Two: Harnessing the wisdom of the crowds. In *Proceedings of the 16th World Wide Web Conference (WWW-07)*, 2007, Banff, Canada, 101-110.

[13] LEE, L. 1999. Measures of Distributional Similarity. In *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics (ACL-99)*, 1999, College Park, Maryland, 25-32.

[14] DAVIDOV, D., RAPPOPORT, A., AND KOPPEL, M. 2007.Fully Unsupervised Discovery of Concept-Specific Relationships by Web Mining. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics(ACL-07)*, 2007, Prague, Czech Republic, 232-239.