REV: Extracting Entity Relations from World Wide Web

Chao Chen  
Department of Computer Science and Technology, East China Normal University  
Shanghai, China  
cchen@ica.stc.sh.cn

Liang He  
Department of Computer Science and Technology, East China Normal University  
Shanghai, China  
lhe@cs.ecnu.edu.cn

Xin Lin*  
Department of Computer Science and Technology, East China Normal University  
Shanghai, China  
xlin@cs.ecnu.edu.cn

ABSTRACT
Quantities of valuable relation knowledge are contained in textual documents on the World Wide Web. However, those data are always organized in semi-structured text and cannot be used directly. We develop an automatic and effective approach to extract relations from World Wide Web, which just requires a few user specified seed instances as input. Those instances are used to generate extraction rules that in turn result in new instances. And in order to improve the reliability of results, an effective method is proposed to assess new extracted instances. This paper introduces the approach in details and the experimental results show that the approach achieves an average precision of 98.67% and can preferably complete the relation extraction task.

Categories and Subject Descriptors
H.3.3. [Information Systems]: Information Storage and Retrieval—information search and retrieval

General Terms
Management, Performance, Human Factors, Verification

Keywords
Entity Relation Extraction, Extraction Rules, Minimum Cover Patterns, Relation Extraction with Verification.

1. INTRODUCTION
Although, quantities of valuable relation knowledge are contained in textual documents on the Web, those data are always organized in semi-structured text and cannot be used directly. For example, to answer the question "who is the mayor of Shanghai, PRC?", a collection of textual documents will provide arguments to the answer Zheng Han is the current mayor of Shanghai. Intuitively, what we should do is finding out those documents and extracting corresponding information. However, this process can be quite time-consuming and laborious for the person who wants to get the knowledge.

Entity Relation Extraction, which can turn unstructured data into structured ones, is proposed to resolve the aforementioned problem. It aims to build an authoritative relation set to rapidly answer relation queries. However, the complexity of natural language makes it quite difficult to solve relation extraction problems. This problem is especially challenging when the sentence structure of the target language is complex and flexible. Firstly, generic patterns used to represent a certain relation are nonexistent or difficult to find. Secondly, it is more difficult to measure the exactness of new instances. So, although much attention has been drawn on this problem, it remains unresolved.

What's more, a lot of approaches that have been proved to be effective on certain small corpus cannot be scaled to big unspecific corpus or unbounded corpus such as World Wide Web.

This paper develops an automatic and effective approach REV (Relation Extraction with Verification) originates from DIPRE [1] to address the entity relation extraction problem. It is based on bootstrapping techniques and barely relies on the lexical and syntactic features of the target language. The only input is a small seed instance set with the form of \(<e_1, e_2, \text{keyword}>\). World Wide Web and search engine are used as data source. Each seed instance is used with the search engine to generate extraction rules. Then, those extraction rules are used to extract new entity instances of the same relation as the seed instance. A method named MCP (Minimum Cover Pattern) is used to assess new instances, which assigns a credibility score to each new instance based on a statistical model of redundancy. An instance with a credibility score higher than the threshold will be used as a seed instance in the next iteration, and then used to construct a structured relation set. This iteration process will be terminated when there is no instance can be extracted. For simplicity, only binary relations are taken into account in this paper and the target language is Chinese. The main contributions of this paper are as follows:

1. This paper develops a new form to represent extraction rules, which aims to get a balance between particularity and universality, so that those extraction rules can capture more new instances and achieve a high precision (Section 3.2).

2. This paper proposes a simple yet efficient strategy named MCP based on a statistical model of redundancy to assess new instances. And a component named MCPA (MCP Assessor) based on it is constructed in the accomplished extraction system, which is used to assess instances and eliminate unreliable instances (Section 3.4).

3. This paper conducts extensive experiments to evaluate the performance of the proposed approach. The experimental results show that our approach performs well under various relations and can preferably complete the entity relation extraction task (Section 4).

The remainder of the paper is organized as follows. Section two introduces related works, including classical extraction
approaches and extraction systems. Section three details REV and the experimental results are reported in section four. Section five concludes the proposed approach followed by a discussion of future work.

2. RELATED WORK
Entity relation extraction is one of the major research fields of information extraction research, and quantities of researchers have made efforts on it. So far, two well-known paradigms have been proposed to address it. Supervised approaches, also called traditional approaches, have been studied for many years. Hand-crafted extraction rules or hand-tagged training data sets are required to construct supervised entity relation extraction systems. Substantial human involvements are also needed in training the system for each new task. Typical systems, such as [2] [3] based on Hidden Markov Models, [4] based on Rule Learning and [5] based on Conditional Random Fields, have been applied to some applications and have difficulties scaling to big corpus. Although supervised approaches have obvious defects, they have been proved to be efficient. [6] [7] get overall F-scores of 73.27% and 76% respectively in Chinese entity relation extraction.

Unsupervised approaches reduce human involvements and eliminate the requirements of predefined relations, hand-taged training data set and hand-crafted rules. In most cases, users make extracting strings between entities in a large corpus as the first step. And then, all the strings gotten from the first step are simplified and clustered to generate relation strings and used as relations of target entity pairs. The advantages of unsupervised approaches are obvious, but those approaches also have inherent drawbacks such as entity pairs of a user specified relation cannot be acquired.

Nowadays, more and more approaches [8] [9] fall into the broad category of bootstrapping techniques and DIPRE is one of the most famous systems. It works in a collection corpus of HTML documents and extracts structured relations. The input of DIPRE is a big corpus, where the tuples to be extracted will tend to appear in uniform contexts repeatedly. DIPRE takes advantages of the redundancy and inherent structure of the corpus to extract the target relation tuples with minimal human involvements.

In this paper, our approach is built on DIPRE, but it develops some key parts of DIPRE. More specifically, REV uses the World Wide Web as data source and uses a new form to represent extraction rules. What’s more, REV exploits an effective strategy to assess instances extracted in each iteration of the extraction process. Only reliable instances will be used as seed instances in the next iteration. We will introduce REV in details in the next section.

3. THE REV SYSTEM
In this section, we will describe REV in details. It is composed of four major components, 1) Sentence generation is responsible for searching for occurrence sentences for relation instances and extraction rules, 2) Extraction rules generation is used for generating extraction rules, 3) New instances are extracted from the new instances generation, and 4) MCP is used to assess new instances. Each component will be introduced in details in following sections.

3.1 Sentence Generation
Sentence generation is used to search for occurrence sentences for relation instances and extraction rules. So, the input of it can be divided into two categories, relation instances and extraction rules.

In REV, relation instances consist of seed instances and extracted instances, in the form of \(<e_1, e_2, \text{keyword}>\). Both \(e_1\) and \(e_2\) represent entities, and \text{keyword} is a short string of several words that denotes the relation between \(e_1\) and \(e_2\). Since REV could extract various kinds of relations, \text{keyword} can denotes every relation we are interested in.

<table>
<thead>
<tr>
<th>(e_1)</th>
<th>(e_2)</th>
<th>\text{keyword}</th>
</tr>
</thead>
<tbody>
<tr>
<td>韩正 (Zheng Han)</td>
<td>上海市 (Shanghai)</td>
<td>市长 (mayor)</td>
</tr>
<tr>
<td>许勤 (Qin Xu)</td>
<td>深圳市 (Shenzhen)</td>
<td>市长 (mayor)</td>
</tr>
</tbody>
</table>

Extraction rules are constructed as 7-dimension tuples and generated in each iteration. And they only have effect in the iteration in which they are generated. The incoming section will introduce extraction rules in more details.

Sentence generation gets the input and generates queries. Then, those queries are sent to search engine. The biggest Chinese search engine [10] is used as the default search engine in REV. Sentence generation retrieves all the segments of text in which every part of a relation instance or an extraction rule occurs effectively.

Sentence generation retrieves all the segments of text in which every part of a relation instance or an extraction rule occurs effectively.

3.2 Extraction Rules Generation
Since the structure of a Chinese sentence is more complex than some other languages, so, in order to make an extraction rule represent the body of a sentence more accurate, REV develops a new form to represent extraction rules and implements a completely automatic component to do this work in the accomplished extraction system. An extraction rule generated from a sentence is a 7-dimension tuple, which can be expressed as one of the following forms:

\(<\text{con}_1, e_1, \text{con}_2, e_2, \text{con}_3, \text{keyword}, \text{con}_4>\)
\(<\text{con}_1, e_1, \text{con}_2, \text{keyword}, \text{con}_3, e_2, \text{con}_4>\)
\(<\text{con}_1, e_2, \text{con}_2, e_1, \text{con}_3, \text{keyword}, \text{con}_4>\)
\(<\text{con}_1, e_2, \text{con}_2, \text{keyword}, \text{con}_3, e_1, \text{con}_4>\)

Figure 1. The main components of REV.
In those six extraction rule forms, $e_1$ and $e_2$ denote two entities respectively; *keyword* means the relation keywords and $con_i$ means a string of several words. It must be noted that $con_i$ just contains words regarded as effective words. We define nouns, verbs and adjectives as effective words in REV. And all the extraction rules must satisfy the following conditions:

$$\sum_{i=1}^{4} \text{length}(con_i) \neq 0$$  \hspace{1cm} (1)

$$\sum_{i=1}^{4} \text{length}(con_i) \leq N$$  \hspace{1cm} (2)

In the above formulas, $\text{length}(con_i)$ means the amount of words that $con_i$ contains. REV sets those two restrictions to get distinguishing results and find a balance between particularity and universality.

It is obvious that all we can get from the sentence generation component is a set of sentences. In the component of extraction rules generation, segmenting words and part-of-speech tagging are the first step. The ICTCLAS [11] is used to do this work, which gets an average precision rate of 98% in Chinese. Then, non-effective words are abandoned. Generally speaking, those potential extraction rules contain too many words and cannot satisfy the conditions introduced above, so the system needs to do some cutting on them. REV uses the cutting rules proposed in [12] to do the work, which are expressed as follows:

$$P_1 = \begin{cases} 
E[L_{\text{min}}; L_{\text{max}}] + E[L_{\text{max}}; L_{\text{rc}}] \\
(\text{if } D_{\text{abs}} = 1 \text{ and } L_s > L_{\text{kw}}; \text{or } L_s < L_{\text{kw}} < L_b) \\
E[L_{\text{lc}}; L_{\text{min}}] + E[L_{\text{min}}; L_{\text{max}}] \\
(\text{if } D_{\text{abs}} = 1 \text{ and } L_b > L_{\text{kw}}) \\
E[L_{\text{lc}}; L_{\text{min}}] + E[L_{\text{min}}; L_{\text{max}}] + E[L_{\text{max}}; L_{\text{rc}}] \\
(\text{other conditions})
\end{cases}$$  \hspace{1cm} (3)

And

$$D_{\text{abs}} = \text{abs}(L_{e1} - L_{e2})$$
$$L_s = \min(L_{e1}, L_{e2})$$
$$L_b = \max(L_{e1}, L_{e2})$$
$$L_{\text{min}} = \min(L_{e1}, L_{e2}, L_{\text{kw}})$$
$$L_{\text{max}} = \max(L_{e1}, L_{e2}, L_{\text{kw}})$$

$L_{\text{lc}}$ and $L_{\text{rc}}$ denote the nearest position of an effective word on the left, right side of the entities or *keyword* respectively, $L_{\text{kw}}$ denotes the position of the *keyword*, $L_{e1}$ and $L_{e2}$ denotes the position of $e_1$ and $e_2$ respectively and $E[m: n]$ means the words between word $m$ and $n$, which includes word $m$.

Table 2. Two extraction rules

<table>
<thead>
<tr>
<th>No</th>
<th>Extraction rules</th>
<th>Content</th>
<th>Extraction rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$e_1$/nr</td>
<td>当选/v</td>
<td>$e_2$/n</td>
</tr>
<tr>
<td>2</td>
<td>$e_1$/ns</td>
<td></td>
<td>$e_2$/n</td>
</tr>
</tbody>
</table>

The extraction rules generated in this component are then sent to the sentence generation. The sentence generation searches for all the sentences that contain a complete extraction rule. After that, those sentences are sent to new instance generation with the corresponding extraction rules, which will be introduced in the coming section 3.3.

### 3.3 New Instance Generation

REV uses the accurate match method to extract new instances. As mentioned above, the sentence generation component just outputs sentences that contain a complete extraction rule. So, each part of a sentence is corresponding to a specific extraction rule.

Segmenting words and part-of-speech tagging are also the first step of extracting new instances. Then, REV cut the sentence into several parts according to the extraction rule that has been used to get this sentence. Finally, REV gets all the potential instances that do not violate entity tags and the extraction rule.

![Image](image.png)

Figure 2. An example of extracting new instances

Table 3. The example of the second extraction rule

<table>
<thead>
<tr>
<th>Position</th>
<th>Content</th>
<th>Extraction rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part_1</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>Part_2</td>
<td>中国/ns</td>
<td>中国/ns</td>
</tr>
<tr>
<td>Part_3</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>Part_4</td>
<td>市长/n</td>
<td>市长/n</td>
</tr>
<tr>
<td>Part_5</td>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>Part_6</td>
<td>韩正/nr</td>
<td>韩正/nr</td>
</tr>
<tr>
<td>Part_7</td>
<td>发表/v</td>
<td>发表/v</td>
</tr>
</tbody>
</table>

Since $e_1$ is at Part 2 and $e_2$ is at Part 6, so we can get two instances from the sentence based on the second extraction rule in table 2. Unfortunately, it is obvious that new instance <中国 (China), 韩正 (Zheng Han), 市长 (mayor)> is not correct, so, filtering all those incorrect instances is quite important for entity relation extraction. REV uses an independent component named MCP to assess potential instances produced by the instance generation. The next section will give a detailed description of assessing potential instances.

### 3.4 MCP

The method used to assess potential instances in REV is MCP (Minimum Cover Pattern) and the component constructed in REV
Because the standard pattern set is generated by the original seed instances, it is not completely authoritative. In order to overcome this drawback, MCPA updates the standard pattern set in each iteration. MCPA collects all the patterns generated by new seed instances that have been assessed to be reliable and uses the following formula to update it.

\[
Occur_{noi,new} = Occur_{noi,old} \times 0.8 + P_i \times 0.2 \quad (7)
\]

\(P_i\) means the occurring number of pattern \(i\) in the newly collected pattern set. \(Occur_{noi,old}\) denotes the occurring number of pattern \(i\) in the old standard pattern set and \(Occur_{noi,new}\) denotes the occurring number of \(i\) in the updated set. As a supplement, MCPA inspects the instance set to filter conflicting instances in each iteration. This step is optional and can be skipped when the relation is not exclusive.

4. EXPERIMENTS

In this section, we will represent the experimental results of REV.

4.1 Precision Measurement Criterion

In this paper, the following formula (8) is used to measure the performance of REV. \(n_{correct}\) means the amount of new instances that are manually assessed to be correct. \(N_{total}\) means the total amount of new instances produced by REV. In this paper, only “市长(mayor)”, “总统(president)” and “部长(minister)” relations are taken into account. Certainly, REV can be used to extract other relations.

\[
\text{Precision} = \frac{n_{correct}}{N_{total}} \quad (8)
\]

4.2 Experimental Results

In this paper, two types of configuration are used to extract new instances in order to measure the performance more accurately. Basic means the basic extraction system without assessing new instances and Complete means the complete extraction system. Table 5 shows the experimental results of three relations in the Basic system. Since the amount of new instances returned by the REV extraction system is always huge, so we randomly select 100 instances from each relation and assess them manually. It needs to be noted that the errors made by ICTCLAS are manually removed before measuring the performance of the proposed approach.

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>市长(mayor)</td>
<td>58%</td>
</tr>
<tr>
<td>部长(minister)</td>
<td>62%</td>
</tr>
<tr>
<td>总统(president)</td>
<td>65%</td>
</tr>
</tbody>
</table>

From table 5, we can find that the Basic system barely get an average precision of 61.67%. This is because that the Basic system does not filter any instance produced by the new instance generation component and all the new instances are reserved. As the example described in 3.3, some of those instances are effective but not correct. Effectivity means that the new instance is satisfied.
with an extraction rule at least and this becomes an inherent defect of the Basic system.

Table 6. The experimental results of three relations in the Complete system

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>Precision</th>
<th>Improvement Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>长（president）</td>
<td>100%</td>
<td>72.41%</td>
</tr>
<tr>
<td>部长（minister）</td>
<td>97%</td>
<td>56.45%</td>
</tr>
<tr>
<td>总统（president）</td>
<td>99%</td>
<td>52.31%</td>
</tr>
</tbody>
</table>

Table 6 shows the experimental results of the complete system. From table 6, we can find that MCPA can eliminate most of the wrong instances and greatly improve the performance of REV. The average improvement rate made by MCPA is 60.39%. However, errors are still exist and cannot be filtered completely. For example, we got a new instance <刘军(Jun Liu), 铁道部 (ministry of railways), 部长(minister)>, which has been assessed to be correct. It cannot be eliminated automatically because it was caused by the journalist who had written a wrong name in the news. And if we just refer to the news it is correct. So, we should do some extra work to reduce the impacts of such errors in the future.

4.3 The Recall Rate of MCPA

The recall rate of MCPA is discussed in details above. The characteristics of the World Wide Web make the precision to be the unique measurement criterion of the REV system. As table 6 shows, MCPA can greatly improve the performance of REV, but in order to measure the performance of MCPA more roundly, we discuss the recall rate of it in this section.

Table 7. The recall rate of MCPA

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>Recall Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>长 (mayor)</td>
<td>78%</td>
</tr>
<tr>
<td>部长 (minister)</td>
<td>85%</td>
</tr>
<tr>
<td>总统 (president)</td>
<td>85%</td>
</tr>
</tbody>
</table>

Table 7 shows the recall rate of MCPA. Through the study of the instances that are correct but have been removed by MCPA, we find that there are two major reasons that cause inaccurate assessments. Firstly, the lack of occurrence sentences of a new specific instance. For example, new instance <刘军(Jun Liu)> just has one occurrence sentence. Secondly, new instances appear in too special patterns. For example, we got twenty patterns for the instance <李志勤(Zhixun Li), 武威市(Wuwei), 市长(mayor)> while only three of those patterns fell into the subset of the standard pattern set. However, since the reliability of relation extraction systems is more important, so the recall rate of MCPA is tolerated and should be improved in our future work.

5. CONCLUSIONS AND FUTURE WORK

This paper proposes an approach REV originates from DIPRE to address the entity relation extraction task. The approach is built on bootstrapping technique and it only needs a few seed instances as input. We also implement a completely automatic relation extraction system based on the approach. Through the experimental results, we can see that the system can get a good performance and preferably complete the relation extraction task.

It needs to be noted that, the approach can be improved in some aspects. In the future, methods used to assess extraction rules are needed. This work is not only important for getting occurrence sentences, but also useful for filtering wrong instances that are effective but not correct. And this work will also improve the efficiency of the relation extraction system because fewer instances are needed to be assessed. Then, a strategy used to improve the recall rate of MCPA while make no influence on its precision is wanted eagerly. Finally, we would like to unify triples extracted by REV into a multi-dimension relation graph to support complex queries.

6. ACKNOWLEDGMENTS

This work is supported by a grant from the Shanghai International Cooperation Foundation (No. 11530700300), two grants from the Shanghai Science and Technology Commission Foundation (No. 11511504000 and No. 10dz50103).

7. REFERENCES