Open Domain Event Attribute Extraction Method

Junlei Zhao, Bicheng Li and Lei Guo

ABSTRACT

To improve the extraction of attribute information related to open domain events, a method using Semantic Dependency Parsing (SDP) and a rule model is proposed. First, a small number of items are extracted from the corpus, and the template library corresponding to the attribute information is summarized. SDP is then used to process the entire corpus, identify unrecognized attribute information templates, and add them to the template library. Finally, the error-driven modification method is used to adjust the generated template library. Experimental results demonstrate that this method can enhance the performance of event attribute information extraction.

INTRODUCTION

In recent years, major events have occurred frequently, both at home and abroad, and the subsequent impact of online public opinion on national stability and social harmony is increasingly important. In the 17-day period after the 5.12 earthquake in Sichuan, Xinhua.com had posted 9294 positive articles, with Chinanews.com and People.com posting 5772 and 3796, respectively. The information contained in these items has a certain regularity and particularity. On the one hand, the attribute information is closely combined with trigger words, and contains a large number of identical descriptions, such as the number of deaths and the number of missing persons; on the other hand, the attribute information is discrete, with long description statements combining both measurable and non-measurable data. There is also unique information such as the hypocenter depth and local rainfall statistics.

1Junlei Zhao, Bicheng Li, Lei Guo, College of Computer Science and Technology, Huaqiao University, Xiamen, China
Obtaining structured attribute information from massive unstructured data has become an urgent problem. In recent years, several researchers have studied the extraction of attribute information. At present, the main techniques include rule models and machine learning methods. In terms of rule model extraction, certain scholars [1][2] have implemented systems that rely solely on a template library to extract text information, reporting relatively good performance. In the machine learning field, human experts [3][4][5] have employed machine learning algorithms to train a corpus, and the resulting trained models have obtained good experimental results. Other researchers [6][7] are using deep learning models to extract attribute information. Although this approach eliminates the need for human participation, current performance levels do not meet practical requirements.

This paper proposes an open domain event attribute information extraction method using Semantic Dependency Parsing (SDP) and an assistant rule model. The structure of this paper is as follows. The next section describes the method proposed in this paper; the third section presents experimental results and analysis; and finally, this study and our research outcomes are summarized in the final section.

OPEN DOMAIN EVENT ATTRIBUTE EXTRACTION METHOD

Principle Analysis and Method Flow

To realize open domain event attribute information extraction, a rule-based event attribute information extraction method is developed. This tool adopts the manual annotation method, extracts the event attribute information expression model set from a randomly selected corpus, and constructs the attribute information extraction rule model.

As the attribute description information of the event exists in open domain information, the current attribute information extraction method does not consider the structure and deep meaning of each sentence, as it is difficult to analyze the relationships among each part of the attribute information in depth. The proposed method analyzes the semantics of corpus texts to obtain the semantic associations between each language unit of the sentence, finds multiple expression models of the same semantics, extracts the semantic models that are not filtered by the rule model, and determines the new attribute information extraction model. This model is added to the model library.

A flowchart of the open domain event attribute information extraction method based on SDP and the rule model described in this paper is shown in Figure 1.
Figure 1. Schematic diagram of attribute information identification and extraction of semantic dependency analysis and rule model.

Construction of Rule Model for Open Domain Event Attribute Information Extraction

Open domain event attribute information contains nouns, noun phrases, or verbs (i.e., trigger words) that characterize the attributes, and attribute values with unit nouns. The attribute values measure or characterize the specific situation of the attribute. Common expressions include “V (verb) + X (attribute value) + (unit noun) + N (attribute noun)”, “N (attribute noun) + X (attribute value) + (unit noun)” and “N (noun phrase) + X (attribute value) + (unit noun)”; the sample collection is presented in Table 1. According to the open domain event attribute information collected, the rule model library is constructed and the open domain corpus text is extracted. A flowchart of the open domain event attribute information extraction based on the rule model is shown by the dotted line in Figure 1.

Attribute Information Extraction Rule Model Supplement Based on SDP

SDP analyzes the semantic association between each language unit of a sentence and presents the semantic association in a dependency structure. The advantage of using semantic dependence to characterize sentence semantics is that there is no need to abstract the vocabulary itself; instead, the vocabulary is described through the semantic framework that the vocabulary constructs. The number of arguments is always a lot smaller than the size of the vocabulary. According to the requirements of the relevant information in the text, the relationship types of some semantic dependency structures are listed in Table I.
In Chinese, different expressions of the same sentence cannot use the same rule model. When various relationship types are extracted, these differences are very subtle and easily overlooked. The number of deaths and the number of injured people are commonly expressed when describing an earthquake event. Because of the different positions of the words, it is impossible to extract the attribute information by using a model, and the difference is not easy to identify. SDP can classify these instances into a class and supplement the rule model library, thus improving the extraction effect.

**Modification of Rule Model Based on Error Driven Modification**

In the process of constructing the rule model, the extracted data are compared with the original text attribute information, and the error between the extracted data and the original text attribute information can be identified. This enables us to find and modify the corresponding rule model. Effectively, the extraction rule model is reset and adjusted. The order is based on the principle of the maximum and longest matching, and is successively decremented to improve the accuracy of attribute information extraction.

**EXPERIMENTAL RESULTS AND ANALYSIS**

**Introduction to the Corpus**

There are 650 flood corpora, 650 typhoon corpora, and 650 earthquake corpora. The data in each corpus is taken from information on webpages, and are mainly based on news events and data collected from major news media such as Xinhua.net, Sina.com, and Sohu.com. The structure of the corpus contains times, places, people, and information about events, in which the time, place, and event are expressed in various forms. Thus, the information descriptions have a wide coverage. In terms of content, the corpora describe events that have occurred in the past ten years. They have strong real-time performance and basically cover most of the major events that have occurred in recent years. From the three types of open domain corpora, 150
papers were randomly selected from each domain as the training corpus for constructing the extraction model and semantic dependency analysis, and the remaining 1500 articles were used as the test corpus.

**Experimental Settings**

The experiment uses the Java version of the HanLP open source tool for data preprocessing. The development environment is Java1.8 and related development kits. The development platform is the eclipse-neon software, with the LTP language processing platform used for semantic analysis. The experimental PC uses an Intel® Core™ i5-3210M CPU @2.50 GHz, 8 GB RAM, and a 64-bit Win10 operating system.

The experimental results are evaluated using the precision P, recall R, and the F1-Measure. The experimental results and performance pairs are presented in Table II.

From the experimental results in Table II, based on a comparison between the rule model (method a) and the proposed technique (method d), it can be seen that the use of semantic dependence analysis significantly improves the F1 value of the event attribute information extraction in all corpora types, indicating that semantic dependency analysis complements the rule model library.

To analyze the overall performance, representative methods developed by Wu et al. [5] (method b) and Yang et al. [8] (method c) were also evaluated. Yang et al. use a weak supervision method to extract attribute information from semi-structured data. Compared with unstructured data, attribute information extraction from semi-structured data has greater operability, so it is important to consider a representative method in the field of semi-structured data in comparative experiments. The method proposed by Wu et al. combines event attribute information extraction with a hidden Markov model and syntactic analysis. This approach is representative of methods that extract attribute information from text corpora in the open domain.

Table II indicates that the extraction of attribute information from the seismic domain by the proposed method is significantly better than that of the two representative methods. The reason for this may be that the original intention of the proposed method is to target open domain unstructured corpora, whereas the weakly supervised learning method is intended for semi-structured corpora such as encyclopedia entries and the method based on a hidden Markov model and syntactic analysis uses statistical machine learning. The recall rate of the latter with seismic attribute information is similar to that of the proposed method, but our method is obviously superior in terms of precision.
TABLE II. EVENT ATTRIBUTE INFORMATION EXTRACTION STATISTICS.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Earthquake attribute</th>
<th>Flood attribute</th>
<th>Typhoon attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P,%</td>
<td>R,%</td>
<td>F,%</td>
</tr>
<tr>
<td>a</td>
<td>88.34</td>
<td>82.13</td>
<td>85.12</td>
</tr>
<tr>
<td>b</td>
<td>79.32</td>
<td>78.49</td>
<td>78.90</td>
</tr>
<tr>
<td>c</td>
<td>85.91</td>
<td>86.39</td>
<td>85.70</td>
</tr>
<tr>
<td>d</td>
<td>90.41</td>
<td>86.60</td>
<td>88.46</td>
</tr>
</tbody>
</table>

CONCLUSIONS

This paper has described an open domain event attribute information extraction method based on semantic dependency analysis and a rule model. To overcome the limitations of existing attribute information extraction methods, semantic analysis is used to analyze the approximate semantics of attribute information expressions, thus improving the extraction performance of open field event attribute information.

At present, open domain event attribute information mining is still in its infancy. In future work, we will investigate extraction methods that can identify attribute information based on techniques mentioned in this paper, such as the method of deep learning used by Zhou and Zhang [9] in comprehensively mining the products of users’ attention. The clustering of attribute information and construction of a reasonable attribute information set will also be examined.

ACKNOWLEDGMENTS

This research is supported by Fujian Social Science Planning Project (FJ2017B073), the Scientific Research Funds of Huaqiao University and Subsidized Project for Postgraduates’ Innovative Fund in Scientific Research of Huaqiao University (1611314020).

REFERENCES