Abstract

XML Topic maps enable multiple, concurrent views of sets of information objects and can be used to different applications. However, to enrich the information or metadata of a topic map or to connect with some document's URI is very labor-intensive and time-consuming. To solve this problem, we propose an approach based on natural language processing techniques to identify and extract useful information in raw Chinese text. Unlike most approaches to parsing sentences based on the integration of complex linguistic information and domain knowledge, we use shallow parsing instead of complex parsing to analyze the constituents and their relationship in discourse. After a document is processed, we may assign this document into a topic node and add the information extracted from the raw text into a topic map.

1. Introduction

The current world-wide web mainly consists of formatted documents, for example, HTML documents, that provide excellent knowledge sources for human consumption. Under the circumstances, computers can only present the formatted results to user rather than process the content of the documents because they lack of the ability of natural language processing. Thus computers are not helpful to encounter the problem of information explosion happened of the current web. An approach employed by the emerging technology, Semantic Web [1,2], is to create a metadata layer that provides semantic descriptions about the content of the formatted documents. Advanced services, for example, conceptual search and semantic navigation can therefore be built upon the machine processable layer.

Topic Maps (XTM) [3] is a XML-based language like RDF [4], are in general used as the carrier of metadata to achieve content interoperability in the metadata layer. A topic map is composed of a number of topics, associations and occurrences [5]. A topic is a reification of subject in the real world. An association indicates the interrelationship between a pair of topics or even more parties. An occurrence connects the information relevant to a subject to the corresponding topic. The metadata can either be created manually using annotation tools [6] or generated automatically by machine. Both approaches have their own merits and disadvantages: the former is laborious, while in the latter it is difficult to build the knowledge bases for processing the texts.

A number of statistical-based methods which can extract a bag of words from documents are used in natural language processing applications, such as information retrieval and text categorization [7,8,9]. These methods may be employed for obtaining topic information in topic maps. However, a bag of words cannot reveal the relationships among these words. In this paper, we propose a method for creating metadata of the topic maps of Chinese documents. The method is to parse each utterance for obtaining the elements and their relationships in texts. For resolving the problem of zero anaphora in Chinese, we employ the method of zero anaphora resolution which had been developed based on the centering model and shallow parsing techniques [10].

In the rest of this paper, we first describe the elements and their relationships in a topic map. In Section 3 we describe in details how to parse Chinese utterances in the shallow level. In Section 4 the method of creation topic maps using the shallow parsing is illustrated. Finally our conclusions are summarized, and future works are suggested.

2. Topic Map

The purpose of a topic map is to convey knowledge about resources through a superimposed layer, or map, of the resources. A topic map captures the subjects of which resources speak, and the relationships between subjects,
in a way that is implementation-independent. The key concepts in topic maps are topics, associations, and occurrences [5]. We now use the examples extracted from [5] to illustrate the relationship among topics, associations, and occurrences:

```
<topic id="hamlet">
  <instanceOf>
    <topicRef xlink:href="#play"/>
  </instanceOf>
  <baseName>
    <baseNameString>Hamlet, Prince of Denmark</baseNameString>
  </baseName>
  <occurrence>
    <instanceOf>
      <topicRef xlink:href="#plain-text-format"/>
    </instanceOf>
    <resourceRef xlink:href="ftp://www.gutenberg.org/pub/gutenberg/etext97/1ws2610.txt"/>
  </occurrence>
</topic>

<association>
  <instanceOf>
    <topicRef xlink:href="#written-by"/>
  </instanceOf>
  <member>
    <roleSpec>
      <topicRef xlink:href="#author"/>
    </roleSpec>
    <topicRef xlink:href="#shakespeare"/>
  </member>
  <member>
    <roleSpec>
      <topicRef xlink:href="#work"/>
    </roleSpec>
    <topicRef xlink:href="#hamlet"/>
  </member>
</association>
```

In the example above, an occurrence containing an addressable information resource, a URL, which is a reference resource of the topic "hamlet". In the example of associations, an association represents the relationship between Shakespeare and the play Hamlet. Because associations express relationships they are inherently multidirectional: If "Hamlet was written by Shakespeare", it automatically follows that "Shakespeare wrote Hamlet"; it is one and the same relationship expressed in slightly different ways. Instead of directionality, associations use roles to distinguish between the various forms of involvement members have in them. Thus the example above may be serialized using natural language as follows: "There exists a 'written by' relationship between Shakespeare (playing the role of 'author') and Hamlet (playing the role of 'work')." Relationships may involve one, two, or more roles [5].

3. Shallow Parsing

Shallow (or partial) parsing which is an inexpensive, fast and reliable method does not deliver full syntactic analysis but is limited to parsing smaller syntactical related constituents [11,12,13,17]. For example, the sentence (1a) and can be divided as (1b).

(1) a. [花蓮 成為 熱門的 旅遊 地點]  
    Hualian become popular NOM tour place        
    Hualien became the popular tourist attraction.

b. [NP 花蓮 ] [VP 成為 ] [NP 熱門的 旅遊 地點]  
    [NP Hualien ] [VP became] [NP the popular tourist attraction]

Given a Chinese sentence, our method of shallow parsing is divided into the following steps: First the sentence is divided into a sequence of POS-tagged words by employing a segmentation program, AUTOTAG, which is a POS tagger developed by CKIP, Academia Sinica. Second the sequence of words is parsed into smaller constituents such as noun phrases and verb phrases with phrase-level parsing. Each phrase is represented as a word list. Then the sequence of word lists is transformed into triples, \([S, P, O]\). For example in (2), (2b) is the output of sentence (2a) produced by AUTOTAG and (2c) is the triple representation.

(2) a. \([\text{Nc} (\text{Nc})(\text{VG})(\text{VH})(\text{DE})(\text{V A}) (\text{Na})]\]

b. \([[[\text{Np}]],[[\text{成為}],[\text{vtp}],[[\text{熱門的,旅遊,地點}],[\text{np}]\]

c. \([[[\text{Np}],[[\text{vtp}],[[\text{熱門的,旅遊,地點}]]]

The definition of triple representation is illustrated below. The triple here is a simple representation which consists of three elements: \(S\), \(P\) and \(O\) which correspond to the Subject (noun phrase), Predicate (verb phrase) and Object (noun phrase) respectively in a clause.

**Definition of Triple**

A triple \(T\) is characterized by a 3-tuple

\[
T = [S, P, O]
\]

where

\(S\) is a list of nouns whose grammatical role is the subject of a clause.

\(P\) is a list of verbs or a preposition whose grammatical role is the predicate of a clause.

\(O\) is a list of nouns whose grammatical role is the object of a clause.

In the step of triple transformation, the sequence of word lists as shown in (2b) is transformed into triples by employing the Triple Rules. The Triple Rules is built by referring to the Chinese syntax. There are four kinds of triples in the Triple Rules, which corresponds to four
basic clauses: subject + transitive verb + object, subject + intransitive verb, subject + coverb\(^1\) + object,\(^2\) and a noun phrase only. The rules listed below are employed in order:

**Triple Rules:**
- \(\text{Triple1}(S,P,O) \rightarrow \text{np}(S), \text{vtp}(P), \text{np}(O)\).
- \(\text{Triple2}(S,P,\text{none}) \rightarrow \text{np}(S), \text{vtp}(P)\).
- \(\text{Triple3}(S,P,\text{none}) \rightarrow \text{np}(S), \text{coverb}(P)\).
- \(\text{Triple4}(S,\text{none},\text{none}) \rightarrow \text{np}(S)\).

The vtp\((P)\) denotes that the predicate is a transitive verb phrase, which contains a transitive verb in the rightmost position in the phrase; likewise the vtp\((P)\) denotes that the predicate is an intransitive verb phrase, which contains an intransitive verb in the rightmost position in the phrase. In the rule \(\text{Triple3}\), the coverb\((P)\) denotes that the predicate is a coverb. The \(\text{Triple4}\) is employed only if an utterance contains only one noun phrase and no other constituent. If all the Triple Rules failed, the ZA Triple Rules are employed to detect ZA candidates.

**ZA Triple Rules**
- \(\text{Triple1}^1(\text{zero},P,\text{zero}) \rightarrow \text{vtp}(P), \text{np}(O)\).
- \(\text{Triple1}^2(S,P,\text{zero}) \rightarrow \text{np}(S), \text{vtp}(P)\).
- \(\text{Triple1}^3(\text{zero},P,\text{zero}) \rightarrow \text{vtp}(P)\).
- \(\text{Triple2}^1(\text{zero},P,\text{none}) \rightarrow \text{vtp}(P)\).
- \(\text{Triple3}^1(\text{zero},P,\text{none}) \rightarrow \text{coverb}(P)\).
- \(\text{Triple4}^1(\text{zero},P,\text{O}) \rightarrow \text{co-conj}(P), \text{np}(O)\).
- \(\text{Triple4}^2(\text{zero},P,\text{O}) \rightarrow \text{prep}(P), \text{np}(O)\).

The zero anaphora in Chinese generally occurs in the topic, subject or object position. The rules \(\text{Triple1}^1\), \(\text{Triple2}^2\) and \(\text{Triple3}^3\) detect the ZAs occurring in the topic or subject position. The rule \(\text{Triple1}^2\) detects the ZAs in the object position and the rule \(\text{Triple1}^3\) detect the ZAs occurring in both subject and object positions. In the rules \(\text{Triple4}^2\) and \(\text{Triple4}^3\), the co-conj\((P)\) and prep\((P)\) denote a coordinating conjunction or a preposition appearing in the initial position of a clause. For example in (3), there are two triples generated. In the second triple, zero denotes a ZA according to \(\text{Triple1}^1\).

(3) a. 張三 參加 比賽 贏得 冠軍 =
    Zhangsan enter competition win champion
    Zhangsan entered a competition and won the champion.

b. [[[張三], [參加], [比賽]], [[[zero], [贏得], [冠軍]]]
    [[[Zhangsan], [enter], [competition]], [[[zero], [win],
    [champion]]]]

After each utterance in text is parsed, we employ the zero resolution method, which was developed in our previous work [10] to resolve zero anaphors in the triples. Then each zero in the triples is replaced by its antecedent. Therefore, the information carried by zero anaphors is also obtained.

## 4. Creation of Metadata in Topic Maps

After each utterance in a document is parsed and transformed into triples, we can not only obtain the relationships between two elements in utterances but further identify the topics in text. The information is used to create the metadata of a topic map. In this section, we first illustrate the creation of the topics and occurrences, and then describe the creation of associations in a topic map.

### 4.1. Topics and Occurrences

One of the most striking characteristics in a topic-prominent language like Chinese is the important element, "topic," in a sentence which can represent what the sentence is about [14]. A topic chain is a frequently used grammatical structure in Chinese occurring in a span of text, where a referent is referred to in the first utterance and the following several utterances talking about the same referent but not overtly mentioning that referent [14]. Grosz et al., in their paper [15], reported on that psychological research and cross-linguistic research have validated that the backward-looking center is preferentially realized by a pronoun in English and by equivalent forms (i.e. zero anaphora) in other languages. By adopting this notion, the key elements of the centering model of local discourse coherence, and the vital characteristic, topic-prominence, in Chinese, we establish the topic identification rule for identifying the topics in text [16].

**Topic identification rule:**

Given grammatical role criteria: Topic > Subject > Object > Others,

For identifying each topic \(t\) in a discourse segment consisting of utterances \(U_1, \ldots, U_m\):

If at least one ZA occurs in \(U_i\),

then choose the antecedent of the ZA as the \(t\) refer to grammatical role criteria

Else if no ZA occurs in \(U_i\),

then choose one element of \(U_i\) as the \(t\) according to grammatical role criteria

End if
In our method, topic chains are used as the source of obtaining occurrences of topics. We therefore employ the topic chain identification method to develop the creation of metadata in topic map. The metadata includes two child elements of the occurrence, resourceRef and resourceData. When the topic chains of a document are identified, we can add either the information of resourceRef to a topic node of a topic map or the information relevant to the topic of the document. For example in (4) has a topic chain, 基隆醫院 ‘Kee-lung hospital’ and we can add an occurrence containing the URL information to the topic 基隆醫院 ‘Kee-lung hospital’.

(4) a. 基隆醫院 ‘扩大 服务 範圍’.
Jilong yiyuan kuoda fuwu fanwei
Kee-lung hospital expand service coverage
Kee-lung General Hospital increases the service coverage.

b. 職業 提升 醫療 服務 品質 及 標準化，
jjji tisheng yiliao fuwu pinzhi ji biaozhunhua
(Kee-lung General Hospital) active improve medical-treatment service quality and standardization
(Kee-lung General Hospital) actively improves the service quality of medical treatment and standardization.

c. 医療 服務 認可 為 辦理 外勞 體檢 醫院。
huo weishengshu renke wei banli wailao tijian yiyuan
(Kee-lung General Hospital) obtain Department-Of-Health certify to-be handle foreign-laborer physical-examination hospital
(Kee-lung General Hospital) is certified by Department of Health as a hospital which can handle physical examinations of foreign laborers.

5. Conclusion
In this paper, we propose a method of creation of metadata using shallow parsing. In addition, we employ the previous work in zero anaphora resolution for obtaining the information omitted in raw text. The method of topic chain identification in Chinese based on the centering model is used as the source of creating occurrences of topics, while the information carried by the triples are used as the source of creating associations.
We have performed that shallow parsing in Chinese text can enrich the information of occurrences, reference resources of topics and associations. We will further work on developing the applications of topic maps, such as information extraction/retrieval system in the future.

6. References