Creation of Topic Map with Shallow Parsing in Chinese

Ching-Long Yeh and Yi-Chun Chen
Department of Computer Science and Engineering
Tatung University
40 Chungshan N. Rd. 3rd. Sec.
Taipei 104 Taiwan
chingyeh@cse.ttu.edu.tw, d8806005@mail.ttu.edu.tw

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 成爲 熱門 的 旅遊 地點 。
Hualian chengwei remen de luyou didian.
Hualian become popular NOM tour place
Hualien became the popular tourist attraction.
b. [NP 花蓮] [VP 成爲] [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. [花蓮(Nc) 成爲(VG) 熱門(VH) 的(DE) 旅遊(VA) 地點(Na)]
b. [[花蓮], np], [[成爲], vtp], [[熱門,的,旅遊,地點], np]
c. [[花蓮], [vtp(成爲)], [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:

Triple1(S,P,O) → np(S), vtp(P), np(O).
Triple2(S,P,*none*) → np(S), vip(P).
Triple3(S,P,*none*) → np(S), coverb(P).
Triple4(S,*none*,*none*) → 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 vip(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 Triple3, the coverb(P) denotes that the predicate is a coverb. The 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

Triple1^{z1}(*zero*,P,O)→ vtp(P), np(O).
Triple1^{z2}(S,P,*zero*)→ np(S), vtp(P).
Triple1^{z3}(*zero*,P,*zero*)→ vtp(P).
Triple2^{z1}(*zero*,P,*none*)→ vip(P).
Triple3^{z1}(*zero*,P,*none*) → coverb(P).
Triple4^{z1}(*zero*,P,O) → co-conj(P), np(O).
Triple4^{z2}(*zero*,P,O) → prep(P), np(O).

The zero anaphora in Chinese generally occurs in the topic, subject or object position. The rules Triple1^{z1}, Triple2^{z1} and Triple3^{z1} detect the ZAs occurring in the topic or subject position. The rule Triple1^{z2} detects the ZAs in the object position and the rule Triple1^{z3} detect the ZAs occurring in both subject and object positions. In the rules Triple4^{z1} and Triple4^{z2}, 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 Triple1^{z1}.

(3) a. 張三 參加 比賽 贏得 冠軍 。
Zhangsan canjia bisai yingde guanjun.
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

[1] A coverb introduces a noun phrase and the phrase formed by the coverb plus the noun phrase generally precedes the main verb and follows the subject or topic [14].

[2] Note that the clause 'subject + coverb + object' is taken from the former part of the syntactic structure containing a coverb: subject/topic + converb + noun phrase + verb + (noun phrase). We leave the later part 'noun phrase + verb + (noun phrase)' from this structure for the rules Triple1 and Trple2 and the rules Triple3 is further employed to establish its ZA Triple rule for detecting the subject/topic omission.

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^{i}擴大 服務 範圍，
Jilong yiyuan kuoda fuwu fanwei
Kee-lung hospital expand service coverage
Kee-lung^{i} General Hospital increases the service coverage.

b. φ_1^i 積極 提升 醫療 服務 品質 及 標準化，
jiji tisheng yiliao fuwu pinzhi ji biaozhunhua
(Kee-lung General Hospital)^{i} active improve medical-treatment service quality and standardization
(Kee-lung General Hospital)^{i} actively improves the service quality of medical treatment and standardization.

c. φ_2^i 獲 衛生署 認可 爲 辦理 外勞體檢醫院。
huo weishengshu renke wei banli wailao tijian yiyuan
(Kee-lung General Hospital)^{i} obtain Department-of-Health certify to-be handle foreign-laborer physical-examination hospital
(Kee-lung General Hospital)^{i} is certified by Department of Health as a hospital which can handle physical examinations of foreign laborers.

```
<topic id="基隆醫院">
  <occurrence>
    <instanceOf>
      <topicRef xlink:href="# plain-text-format "/>
    </instanceOf>
    <resourceRef  xlink:href=
     "URL_Of_The_News_About_基隆醫院"/>
  </occurrence>
</topic>
```

4.2. Associations

A triple is characterized by a 3-tuple, [*S*, *P*, *O*], where *P* denoting the predicate states the connection between the other elements, *S* and *O*. Because associations in topic maps express relationships between topics, we add the information carried by a triple into a topic map as an association if *S* and *O* both match existing topics. For example in (2c), if both 花蓮 'Hualian' and 旅遊地點 'tourist attraction' are two existing topic, we can add an association in to a topic map.

```
<association>
  <instanceOf><topicRef xlink:href="#成爲"/>
  </instanceOf>
  <member>
    <roleSpec><topicRef xlink:href="#place_name"/>
    </roleSpec>
    <topicRef xlink:href="#花蓮"/>
  </member>
  <member>
    <roleSpec><topicRef xlink:href="#location"/>
    </roleSpec>
    <topicRef xlink:href="#旅遊地點"/>
  </member>
</association>
```

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

[1] Cost, R. Scott et al., ITtalks: a case study in the Semantic Web and DAML+OIL, *IEEE Intelligent Systems Special Issue*, 2002.

[2] Staab, S. et al., Semantic community web portals, WWW9, Amsterdam, 2000.

[3] Pepper, Steve and Moore, Graham. ed., 2001, XML Topic Maps (XTM) 1.0., *TopicMaps.Org Specification*.

[4] Lassila, O. and Swick, R. R. ed., 1999, Resource Description Framework (RDF) Model and Syntax Specification, *W3C Recommendation*.

[5] Biezunski, Michel, Bryan, Martin, and Newcomb, Steven R., ed., 1999, *ISO/IEC 13250 Topic Maps: Information Technology -- Document Description and Markup Languages*.

[6] Benjamins, R., Fensel, D., Decker, S., and Gomez, A., 1999, KA2: building ontologies for the Internet: a mid term report, *International Journal of Human Computer Studies*, pp. 687-712.

[7] Salton, G. and Buckley, C., 1988, Term Weighting Approaches in Automatic Text Retrieval, Information Processing and Management, 24(5):513-523.

[8] Schapire, R., Singer, Y. and Singhal, A., 1998, Boosting and rocchio applied to text filtering, Proceedings of SIGIR-98, 21st

ACM International Conference on Research and Development in Information Retrieval, pages 215-223, Melbourne, Australia.
[9] Tsay, Jyh-Jong and Wang, Jing-Doo, 2000, Design and Evaluation of Approaches to Automatic Chinese Text Categorization, Computational Linguistics and Chinese Language Processing (CLCLP), 5(2): 43-58.
[10] Yeh, Ching-Long and Chen, Yi-Chun. 2003. Zero anaphora resolution in Chinese with partial parsing based on centering theory. In *Proceedings of IEEE NLP-KE03*, Beijing, China.
[11] Abney, Steven, 1991, Parsing by chunks, in: Robert Berwick, Steven Abney, and Carol Tenny, editors, *Principle-Based Parsing*, Kluwer Academic Publishers.
[12] Li, X. and Roth, D., 2001, Exploring Evidence for Shallow Parsing, *Proceedings of Workshop on Computational Natural Language Learning*, Toulouse, France.
[13] Mitkov, Ruslan, 1999, Anaphora resolution: the state of the art, *Working paper* (Based on the COLING'98/ACL'98 tutorial on anaphora resolution), University of Wolverhampton, Wolverhampton.
[14] Li, Charles N. and Thompson, Sandra A., 1981, *Mandarin Chinese – A Functional Reference Grammar*, University of California Press.
[15] Grosz, B. J., Joshi, A. K. and Weinstein, S., 1995, Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics,* 21(2), pp. 203-225.
[16] Yeh, Ching-Long and Chen, Yi-Chun, 2004, Topic Identification in Chinese Based on Centering Model. *Proceedings of ACL Workshop on Reference Resolution and Its Applications*, Barcelona, Spain.
[17] Abney, Steven, 1996, Tagging and Partial Parsing. In: Ken Church, Steve Young, and Gerrit Bloothooft (eds.), *Corpus-Based Methods in Language and Speech*. An ELSNET volume. Kluwer Academic Publishers, Dordrecht.
[18] Kuang-Hua Chen, 1995, Topic Identification in Discourse, *Proceedings of EACL 1995*, pp. 267-271.
[19] Tadashi Nomoto and Yuji Matsumoto, 1996, Exploiting Text Structure for Topic Identification, Proceedings of the 4th Workshop on Very Large Corpora, pp.101-112.