Creation of Topic Map by Identifying Topic Chain in Chinese

Ching-Long Yeh and Yi-Chun Chen
Department of Computer Science and Engineering
Tatung University
40 Chungshan N. Rd. 3rd. Section
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. For example, thesaurus-like interfaces to corpora, navigational tools for cross-references or citation systems, information filtering or delivering depending on user profiles, etc. However, to enrich the information 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 traditional approaches to parsing sentences based on the integration of complex linguistic information and domain knowledge, we work on the output of a part-of-speech tagger and use shallow parsing instead of complex parsing to identify the topics of sentences. The key elements of the centering model of local discourse coherence are employed to extract structures of discourse segments. We use the local discourse structure to solve the problem of zero anaphora in Chinese and then identify the topic which is the most salient element in a sentence. After we obtain all the topics of a document, we may assign this document into a topic node of the topic map and add the information of the document into the topic element simultaneously.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.7 [**Document and Text Processing**]: Miscellaneous

General Terms
Design, Experimentation.

Keywords
Topic Maps, Topic Identification, Shallow Parsing, Zero Anaphora Resolution, Centering Model.

DocEng'04, October 28–30, 2004, Milwaukee, Wisconsin, USA.

1. INTRODUCTION
The current world-wide web mainly consists of formatted documents, for example, HTML documents, that provide 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], 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.

In general, XML-based languages, for example RDF, Topic Maps (XTM) [2], are used as the carrier of metadata to achieve content interoperability in the metadata layer. The metadata can either be created manually using annotation tools 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. In this paper, we propose a cheaper while reliable method for the automatic creation of metadata. We aim at creating the topic maps of Chinese documents by recovering the topics in the texts.

A topic map is composed of a number of topics, associations and occurrences [3]. 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 granularity of occurrence can range from a whole document to a span of text in document. In Chinese text, a topic usually occurs in the initial position of an utterance. A topic chain occurs in a span of text, where the topic of the beginning occurrence is referred by the succeeding utterances whose topics are omitted, termed zero anaphor. We have developed a method of zero anaphora resolution based on centering model and shallow parsing techniques [4]. In this paper, we employ the method to identify the topic chains in a text. Observing Chinese written texts, the span of text covered by a topic chain to a large extent represents the information about a certain subject. In other words, a topic chain can be 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.

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* [3]. We now use the examples extracted from [3] 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/etext97/1ws2
          610.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, an 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 [3].

3. TOPIC CHAIN IDENTIFICATION

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 [5]. That is, if we can identify topic chains from Chinese sentences, we can obtain the most information embedded in the text. In this paper, we tend to identify the topic chains within a discourse based on the centering model [6] and use the techniques of shallow parsing [7].

3.1 Centering Model

In the centering theory [6], the 'attentional state' was identified as a basic component of discourse structure that consisted of two levels of focusing: global and local. For Grosz and Sidner, the centering theory provided a model for monitoring local focus and yielded the centering model which was designed to account for the difference in the perceived coherence of discourses. In the centering model, each utterance U in a discourse segment has two structures associated with it, called forward-looking centers, $C_f(U)$, and backward-looking center, $C_b(U)$. The forward-looking centers of U_n, $C_f(U_n)$, depend only on the expressions that constitute that utterance. They are not constrained by features of any previous utterance in the discourse segment (DS), and the elements of $C_f(U_n)$ are partially ordered to reflect relative prominence in U_n. Backward-looking centers, C_bs, are often omitted or pronominalized and discourses that continue centering the same entity are more coherent than those that shift from one center to another. Grosz *et al*., in their paper [6], assume that grammatical roles are the major determinant for ranking the forward-looking centers, with the order "*Subject* > *Object(s)* > *Others*". The superlative element of $C_f(U_n)$ may become the C_b of the following utterance, $C_b(U_{n+1})$.

In addition to the structures for centers, C_b, and C_f, the centering model specifies a set of constraints and rules [6].

3.2 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 constituents such as noun phrases or verb phrases [7,8]. For example, the sentence (1a) can be divided as (1b):

(1) 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. [NP 張三] [VP 參加] [NP 比賽] [VP 贏得] [NP 冠軍]
[NP Zhangsan] [VP enter] [NP competition] [VP win] [NP champion]

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

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*].

The triple consists of three elements: *S*, *P* and *O* which correspond to the *Subject*, *Predicate* and *Object* respectively in a clause. *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. There are four kinds of triples, which corresponds to four basic clauses: subject + transitive verb + object, subject + intransitive verb, subject + preposition + object, and a noun phrase only. If topic, subject or object is omitted in these clauses, the zero anaphora occurs.

For example in (1), there are two *triples* generated. In the second *triple* of (1c), *zero* denotes a zero anaphor.

3.3 Topic Identification

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 [5].

Topic identification is similar to theme identification in [8]. The key elements of the centering theory, forward-looking centers and backward-looking center are employed to identify themes. The theme clearly corresponds to the backward-looking center: the theme, under a general definition, is what the current utterance is about; what utterances are about provides a link to previous discourse, since otherwise the text would be incoherent. The role of the backward-looking center is precisely to provide such a link. In our approach to topic identification, we employ the topic identification rule based on centering model to identify the topic. When a ZA occurs in the utterance U_i, the antecedent of the ZA is identified as the topic of U_i. Otherwise, if the transition relation, center shifting, occurs, topic will not be identified as any of the element in the preceding utterance but the element in the current utterance according to grammatical role criteria. The topic identification rule is described below:

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, … , 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

4. CREATION OF METADATA

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 (2) has a topic chain, 基隆醫院 and we can add an occurrence containing the URL information to the topic 基隆醫院.

(2) a. 基隆醫院^i 爲 擴大 服務 範圍，

Jilong yiyuan wei kuoda fuwu fanwei

Kee-lung hospital for expand service coverage

Kee-lung^i General Hospital aims to increase 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 為 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>
```

5. CONCLUSION

Based on observations on real texts, we found that to identify the topics in Chinese context is much related to the issue of zero anaphora resolution. We use a zero anaphora resolution method to resolve the problem of ellipsis in Chinese text. The zero anaphora resolution method works on the output of a part-of-speech tagger and employs a shallow parsing instead of a complex parsing to resolve zero anaphors in Chinese text.

In this paper, we propose a method of creation of metadata by topic chain identification in Chinese based on the centering model. The span of text covered by a topic chain to a large extent represents the information about a certain subject. That is, topic chain can be used as the source of obtaining occurrences of topics. We have performed that topic chain identification can either enrich the information of occurrences or obtain reference resources of topics. We will further work on the extraction of other metadata like *association* and develop the applications of topic maps, such as information extraction/retrieval system in the future.

6. ACKNOWLEDGMENTS

We give our special thanks to CKIP, Academia Sinica for making great efforts in computational linguistics and sharing the Autotag program to academic research.

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

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

[4] 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.

[5] Li, Charles N. and Thompson, Sandra A. 1981. *Mandarin Chinese – A Functional Reference Grammar*, University of California Press.

[6] 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.

[7] 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.

[8] Rambow, O. (1993). Pragmatic aspects of scrambling and topicalization in German: A Centering Approach. In *IRCS Workshop on Centering in Discourse*. Univ. of Pennsylvania, 1993.