RDF, Metadata, Information Retrieval

Cecilia Wong
City University of Hong Kong

Abstract

In addition to the typical methods employed in information retrieval systems, e.g. calculating frequency of keywords, pattern matching involving keywords, I am proposing an approach to information search and retrieval based not only on the basic element set known as the Dublin Core Metadata Element Set (DCMES), but also based on the identification of linguistic information about the rhetorical structure of the text. This rhetorical structure information may be inferred from linguistic cues identified and tagged using RDF. The cues and criteria in identifying rhetorical structure information are based on those developed by Corston-Oliver(1998).

The text base in question consists of abstracts of linguistics journal articles drawn from a collection of over three hundred papers on the topic of Chinese Linguistics. Included in our text base are abstracts from linguistics journals in both Chinese and English. Information retrieval is web-based. Besides offering a search and retrieval capability, I also am developing a web interface for authors or publishers to submit their abstracts to the text base.

■ **Approximate Word Counts: 2193**

Introduction

In addition to the typical methods employed in information retrieval systems, e.g. calculating frequency of keywords, pattern matching involving keywords, I am proposing an approach to information search and retrieval based not only on the basic element set known as the Dublin Core Metadata Element Set (DCMES), but also based on the identification of linguistic information about the rhetorical structure of the text. This rhetorical structure information may be

inferred from linguistic clues identified and tagged using RDF. The cues and criteria in identifying rhetorical structure information are based on those developed by Corston-Oliver(1998).

The text base in question consists of abstracts of linguistics journal articles drawn from a collection of over three hundred papers on the topic of Chinese Linguistics. Included in our text base are abstracts from linguistics journals in both Chinese and English. Information retrieval is web-based. Besides offering a search and retrieval capability, we also are developing a web interface for authors or publishers to submit their abstracts to the text base.

What is an abstract? According to Kelly (1990:14), an abstract "should state: [1] what the Report is concerned with; [2] the theoretical framework used; [3] the method of investigation; [4] institutions who participated if any; [5] major findings; and [6]conclusions and implications of the findings."

■ Describing the texts using Dublin Core

Dublin Core provides the underlying metadata framework for describing our collection of abstracts in terms of Title , Creator, Subject, etc. The advantage of Dublin Core is that it uses "a common vocabulary for classifying information" (Laurent and Biggar, 1999, p. 225). This vocabulary is basically divided into three aspects, including Content, Intellectual Property and Instantiation. It attempts to serve as a framework for interdisciplinary information, so that interoperability can be achieved. Moreover, Dublin Core can be easily extended using two types of extensions, one for refining or enhancing the meaning of elements, another for refining or enhancing the interpretation of values (Miller, Miller & Brickley, 1999).

According to Miller and Weibel ("Metadata With a Mission: Dublin Core" published in XML.com Oct. 25 2000), "members of the RSS community have recently been advocating RDF as a powerful, modular means of combining semantics defined by Dublin Core with additional vocabularies (syndication, aggregation, threading) to produce effective site summaries and syndication services". I am similarly employing RDF as a means of combining the semantics of Dublin Core with a linguistically-rich vocabulary designed to enable

linguistic exploration of the rhetorical structure of the abstracts, thereby rendering a query capability which draws on a probable interpretation of the texts in question.

■ Describing the linguistic clues of rhetorical structure

Rhetorical Structure Theory (RST) was developed at USC Information Sciences Institute by William C Mann and Sandra A Thompson. RST is primarily aimed at describing those functions and structures that make texts effective and comprehensible tools for human communication (Mann, et.al. 1992:43). Based on their observations of texts, Mann and Thompson concluded that considerations of text structure concern pairs of regions of text, which are related in some way. Rhetorical Structure Theory (RST) aims to describe natural texts, "characterizing their structure primarily in terms of relations that hold between parts of the text" (1987:1). The following summarizes the underlying assumptions of RST (Mann et.al. 1992:43-46):

- Organization. Texts consist of functionally significant parts.
- Unity and coherence. There must be sense of unity to which every part contributes. Unity and coherence arise from imputed function. Text is perceived as having unity and coherence because all its parts are seen as contributing to a single purpose of the writer (created to achieve a single result).
- Hierarchy. Elementary parts of a text are composed into larger parts, which in turn are composed of yet larger parts up to the scale of the text as a whole.
- Homogeneity of hierarchy. The same functional description applies at every scale.
- Relational composition. A small, finite set of highly recurrent relations holding between pairs of parts of text is used to link parts together to form larger parts.
- Asymmetry of relations. The most common type of text structuring relation is an asymmetric class, called nucleus-satellite relations. One member of the pair is more central (nucleus), the other more peripheral (satellite).

Using RST for text analysis, clauses or text spans are classified as being nuclei or satellites, clearly identifying which information is considered by the analyst to be more central. Also, the relations among the clauses or text spans are identified using RST in order to illustrate the development of a coherent text. By identifying the different relations in a text, one can better ascertain how the various spans in a text combine to form a coherent and meaningful text.

RST offers a systematic approach to interpretation of textual meaning. A recent attempt at automatic RST analysis (Corston-Oliver (1998)) has illustrated how linguistic cues, e.g. voice, grammatical dependency relations and conjunctions, may be used to determine the rhetorical structure. For example, the use of consecutive conjunctions or adverbials, like firstly, secondly, next, then, finally, indicate a sequential structure. Besides, there are criteria to be fulfilled in order to determine whether the text spans belong to a certain relation as the same linguistic cues can be used to indicate more than one relation. Corston Oliver also introduced a heuristic scoring procedure to aid in the determination of text relations. For example, if text spans contain the subordinate conjunction "whereas", 30 points would be assigned to indicate the likelihood of the AsymmetricContrast relation between text spans. I intend to tag these linguistic cues and criteria in the texts as the basis for subsequent exploration of rhetorical relations using the inferencing capability of RDF.

The rhetorical structure of texts also plays a significant role in facilitating search and retrieval. By identifying the nuclei and satellites of different text spans in texts, the core information of the texts can be differentiated from the more peripheral information. Mann and Thompson point out that if though all the satellite spans were deleted, the remainder would still be a coherent text (1988:267). Since the contents of nuclear spans are the major concern of a text, it suggests the likelihood that keywords occurring only in satellite spans in a text may not be among the major findings or conclusions of the paper. Corston-Oliver suggests that the retrieval result from a statistical approach can be improved by weighting in favor of the nuclei spans (1999:238). Through the process in determining the rhetorical

relations among text spans, I intend to make the nuclear spans identifiable.

Certain rhetorical relations appear more prominently in abstracts, e.g. solutionhood and interpretation. Typically, the nuclear span in an interpretation relationship highlights major findings and the conclusion. On the other hand, the nucleus in a solutionhood relationship may refer to the method of investigation, while the satellite(s) may refer to the theoretical framework. Our goal is to ascertain the kind of relations operating between text spans through a kind of linguistic exploration based on encoded linguistic cues. Exploration into possible rhetorical relations between text spans together with the identification of nuclear spans will enable search and retrieval of information along the lines of such queries as: 'retrieve all abstracts whose theoretical approach is based on optimality theory'; 'retrieve all abstracts whose methodology is qualitative and/or ethnographic'. However, these prominently relations, solutionhood and interpretation are among those relations lacking in Corston-Oliver's. I propose to add these to his list as well as define a set of cues and criteria for Chinese data.

■ Linguistic exploration through RDF

In my RDF schema, there are two kinds of information tagged, including bibliographical ones, such as, abstract number, author's name dedicated by Dublin Core and textual information, like voice and grammatical dependency. In this paper, I will concentrate on discussing the textual one. Linguistic information, such as, voice, grammatical dependency and conjunction are represented as attributes in the RDF schema. An example is illustrated below:

```
<rdf:Property rdf:about="voice">
<SLOT-MAXIMUM-CARDINALITY>1</SLOT-MAXIMUM-
CARDINALITY>
<SLOT-VALUE-TYPE>Active</SLOT-VALUE-TYPE>
<SLOT-VALUE-TYPE>Passive</SLOT-VALUE-TYPE>
<SlotValueTypeExtension>Symbol</SlotValueTypeExtension>
<domain rdf:resource="Clause"/>
</rdf:Property>
```

An inference tool constructed in SWI-Prolog, developed by Geoff Chappell, is used to help achieve improved results from user queries. For example, it may be inferred from the presence of those conjunctions or adverbials typically showing sequential structure that certain text spans cover aspects of the research methodology.

Before entering the inferencing stage, RDF statements must first be parsed using the SWI-Prolog RDF parser created by Jan Wielemaker. Parsing is needed to preprocess the RDF specification into a list of rdf (Subject, Predicate, Object) triplets, which may be more easily manipulated and navigated during the inferencing stage. Inferences can be drawn from the two different kinds of tagged information. Illustrated below is an example of textual information. The following assertions are in the RDF model after parsing:

```
rdf('http://www.rdfschema.org/mynamespace.rdf#clause10_3',
rdf:conjunction, literal('on the other hand')).
rdf('http://www.rdfschema.org/mynamespace.rdf#clause10_3',
rdf:dependency, literal('main')).
rdf('http://www.rdfschema.org/mynamespace.rdf#clause10_3',
rdf:voice, literal('passive')).
```

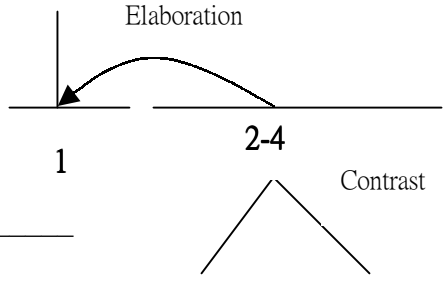
The criteria and cues developed by Corston-Oliver are transformed into rules in Prolog so that the RDF data can be checked with the rules in order to determine if they belong to certain relations. More specifically, rules of criteria will be checked first. Text spans are then determined if they belong to certain relations. Then, rules of cues will be processed to check whether the text spans contain specific cues. Cue numbers will then be assigned to those text spans. Scores adopted from Corston-Oliver's heuristic scoring will be given to the text spans according to the cue numbers. After these processes are done, different text spans may belong to more than one possible relation while some belong to none. The values of score are then used to arrange the relations into a list from higher marks to lower ones. Full stops or other sentence final punctuations will be used to group different text spans into sentences. Sentences with more than one text spans will be considered in advance.

I proposed a sequence of procedures to construct the RST structure of the data. While considering a sentence, the first thing to be checked is

whether there is more than one relation belong to the sentence (see Figure 1). If so, as in text spans 3 and 4, the relation of the last text span with the highest mark will be used to be the relation between the spans. In this case, Result will be the relation between text spans 3 and 4. The text span originally contains the relation will be the nuclear of the relation, i.e. text span 4. For the text spans in front of them, their belonging relations will be checked. As in this example, text spans 1 and 2 got no relation. Thus, two empty text spans are found. Relation from the next text span that contains relation will be extracted to take their places. The amount of relations being extracted will be the same as the empty text spans found. In this case, the two highest marks relations will be taken out, i.e. Elaboration and Contrast. The relation with higher marks will be the relation in a higher level of the hierarchy. Hence, Contrast will be used to link up text span 2 with 3 and 4 followed by linking up text span 1 with that 2 to 4 span using Elaboration. The newly added text span will then become the nuclear of the relation, i.e. the second text span and then the first one. Since the Contrast relation is a symmetric relation, text span 2 and 3-4 are both nuclei for the relation. The resultant structure is as shown in Appendix 1: Graph 1. There are two slots labeled as X and Y as no indication of the relations can be found. I compared this result with an analysis result done manually and found that they are perfectly compatible except for the empty slots labeled X and Y here. They coincidentally are the slots labeled Solutionhood and Interpretation in the manual analyzed result. Therefore, I further prove that these two missing relations in Corston-Oliver's list are important in such kind of academic abstract data. I planned to define cues and criteria of these two relations and add to his list.

List of relations found in Abstract Number 10:
A stochastic finite-state word-segmentation algorithm for Chinese
(Sproat R., Gale W., Shih C., & Chang N., 1996).

(The dividing straight line indicating sentence division)



Text span (1):

Text span (2):

Text span (3): Elaboration -> Contrast -> Circumstance
Text span (4): Result -> Purpose

Text span (5):

Text span (6): Elaboration
Text span (7): Cause -> Elaboration -> Contrast -> Sequence -> Purpose

Text span (8):
Text span (9): Purpose

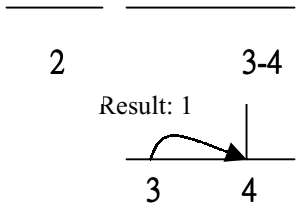


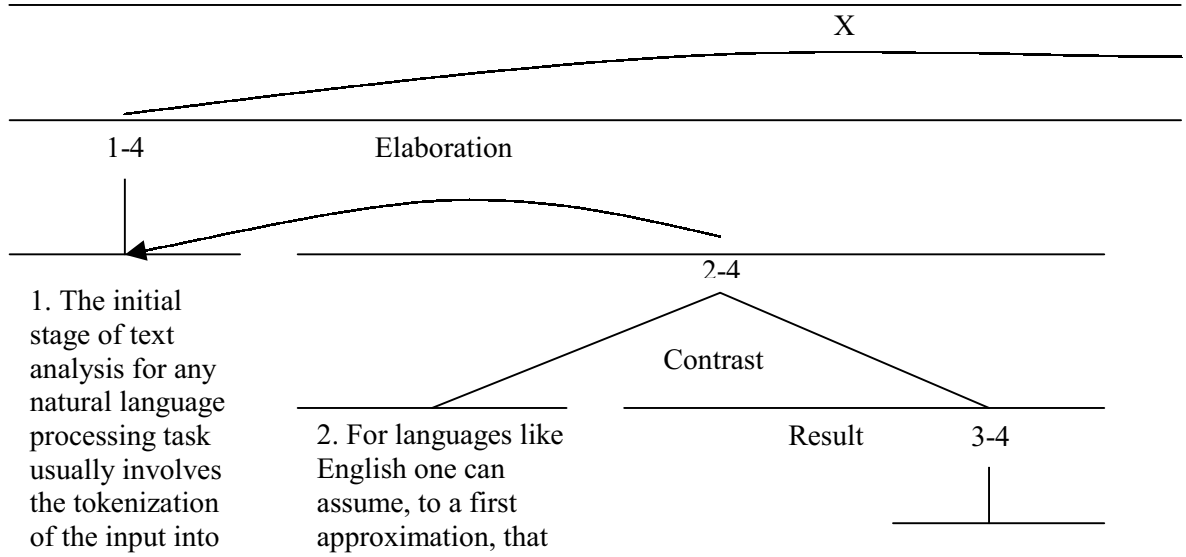
Figure 1

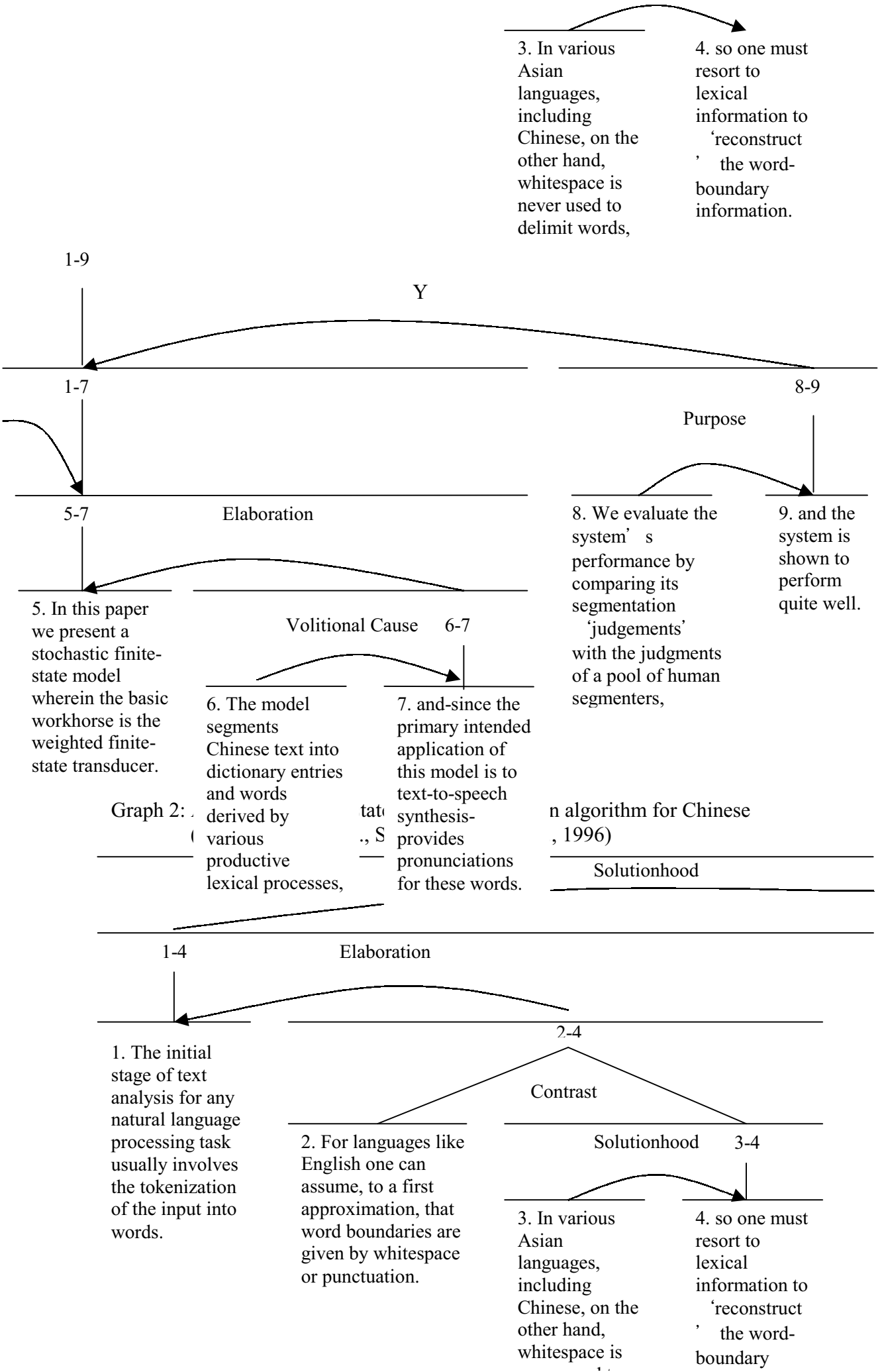
Conclusion

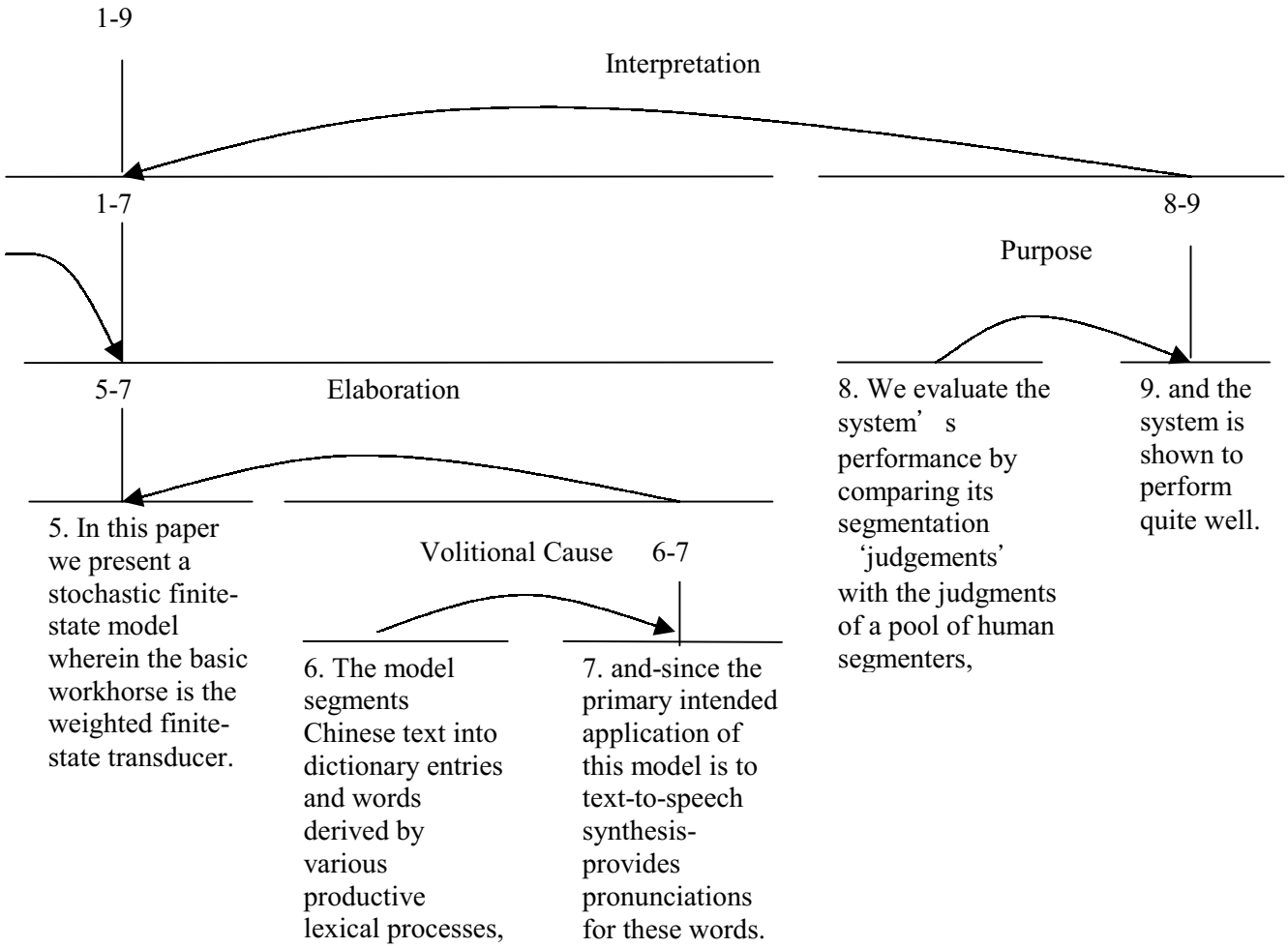
Tim, Berners-Lee (1998) describes the Semantic Web as "a web of data, in some ways like a global database." RDF makes it possible to declare a knowledge base which may be further extended through inferencing. In many ways, RDF brings together the advantages of an object database and the programming power of a logic programming language like Prolog. Using RDF, linguistic data may be encoding in a machine-readable format from which inferences can be drawn by machine about the structure and meaning of texts. Automatic of the process can bring not only efficiency but also effectiveness to the approach.

Appendix I: graph 1:result drawn from program, graph 2:manual analysis.

Graph 1: A stochastic finite-state word-segmentation algorithm for Chinese
(Sproat R., Gale W., Shih C., & Chang N., 1996)




3. In various Asian languages, including Chinese, on the other hand, whitespace is never used to delimit words,
4. so one must resort to lexical information to 'reconstruct' the word-boundary information.
1-9
Y
1-7
8-9
Purpose
5-7
Elaboration
8. We evaluate the system's performance by comparing its segmentation 'judgements' with the judgments of a pool of human segmenters,
9. and the system is shown to perform quite well.
5. In this paper we present a stochastic finite-state model wherein the basic workhorse is the weighted finite-state transducer.
Volitional Cause
6-7
6. The model segments Chinese text into dictionary entries and words derived by various productive lexical processes,
7. and-since the primary intended application of this model is to text-to-speech synthesis-provides pronunciations for these words.
Graph 2:
tat
., S
n algorithm for Chinese
, 1996)
Solutionhood
1-4
Elaboration
1. The initial stage of text analysis for any natural language processing task usually involves the tokenization of the input into words.
2-4
Contrast
2. For languages like English one can assume, to a first approximation, that word boundaries are given by whitespace or punctuation.
Solutionhood
3-4
3. In various Asian languages, including Chinese, on the other hand, whitespace is
4. so one must resort to lexical information to 'reconstruct' the word-boundary



References

Berners-Lee, T. (1998.10.14). Semantic Web Road map [On-line]. Available HTTP: http://www.w3.org/DesignIssues/Semantic.html

Corston-Oliver, S. (1998). Computing representations of the structure of written discourse. U.S.A.: UMI Company.

Kelly, M. (1990). Writing Your Project Report - Workshop Series No. 1. Hong Kong: Educational Technology Centre, City Polytechnic of Hong Kong.

Laurent, S. & Biggar, R. (1999). Organizing information: RDF and Dublin Core. In Inside XML DTDs. U.S.A.:McGraw-Hill Companies Ltd.

Mann, W. (1999.11.23). The Two Frameworks Text [On-line]. Available HTTP: http://www.sil.org/linguistics/rst/2framewk/index.htm

Mann, W., & Matthiessen, C. (1991.12). Functions of language in two frameworks. Word, 42(3), 231-249.
Mann, W., Matthiessen, C., & Thompson S. (1992). Rhetorical Structure Theory and text analysis. In Mann, C., & Thompson, S. (eds) Discourse Description: Diverse linguistics analyses of a fund-raising text. U.S.A.: John Benjamins Publishing Co.
Mann, W., & Thompson, S. (1988). Rhetorical structure theory: Toward a functional theory of text organization. Text, 8(3), 243-281.
Miller, & Weibel (2000.10.25). Metadata With a Mission: Dublin Core [On-line]. Available HTTP: http://www.xml.com
Miller, E., Miller, P., & Brickley, D. (1999.7.1). Guidance on expressing the Dublin Core within the Resource Description Framework (RDF) [On-line]. Available HTTP: http://www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/
Thompson, S., & Mann, W. (1987.7). Rhetorical Structure Theory: A framework for the analysis of texts. IPrA-Papers-in-Pragmatics, 1(1), 79-105.