Proceedings of the First Asia Digital Library Workshop
6-7 August 1998, Hong Kong, pp. 123-135.

Digital Libraries, Metadata and Text Processing

Kuang-hua Chen

Department of Library and Information Science

National Taiwan University

Taipei, TAIWAN

khchen@steelman.ls.ntu.edu.tw

Abstract

Accompanying fast development of the Internet, the concept of digital libraries is widely accepted and discussed by researchers. The experiences of physical libraries say that it is necessary to apply some means like bibliographic control to fulfill high-quality Internet services. Metadata is a key approach based on this line. From the computer scientists' viewpoint, information extraction is a similar task, which extracts the appropriate information based on predefined templates but with automatic procedures. How to integrate metadata and information extraction with techniques of text processing will be an important issue for digital libraries. This paper discusses the relationship between metadata and information extraction and proposes a fast parsing algorithm for text processing. Experimental results are also reported.

1. Introduction

The Internet is quickly growing and the resources around the cyberspace are also quickly accumulating. However, a great deal of garbage data confuse users and make them get lost. How to distil the proper information becomes an indispensable issue. Under this circumstance, many researchers devote themselves to trying to provide possible solutions. On the one hand, library scientists think that the data in the Internet as a collection of digital objects. Naturally, an extension of physical library to digital library is widely discussed. On the other hand, computer scientists manage to take care of this issue from an extension of information retrieval system to Internet. Nevertheless, both sides agree that a new concept of “digital library” is clearly recognized. However, due to the diversity of digital libraries, how to precisely define digital libraries remains controversy. We don’t give a definition of digital library here. In stead, we cite the words of National Science Foundation (NSF) to describe the function of digital libraries. That is, digital libraries provide “the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval and processing via communication networks -- all in user-friendly ways”. [NSF/ARPA/NASA, 1992]

The purpose of researches of Information Retrieval (IR) is to help human to find out the useful information. Some Internet service-providers allow people to use their search engines to search for full-text indexed documents. However, the results cannot fulfill what people expect. On the contrary, the readers of libraries access the appropriate collections via OPAC (Online Public Access Cataloging) or WebPAC

(Web-based Public Access Cataloging), and the search results usually attain the users' information need. The reason is that the resources of libraries are carefully cataloged based on some kind of well-defined metadata format, but the resources of the Internet are not. Bibliographic control on libraries' collection has been used for a very long period. Library scientists have many useful experiences of organizing information and know how to effectively maintain information. It is the high time to apply these experiences to organizing and maintaining Internet resources.

Metadata is used to describe entities including people, document, book, map, and so on. A high-quality and appropriate metadata can effectively characterize entities and easily distinguish entity from entity. As a result, to search a specific entity becomes a much easier task. However, the variance of entities makes situation more complicated. For treasure collection such as historic objects, it is reasonable to describe it more detailed. Therefore, CIMI metadata format [CIMI, 1997] and THA metadata format [Chen, 1998] are complicated. In contrast, the Dublin Core is simple in principle. From this viewpoint, minimalists and structuralists are two sides of a spectrum of metadata types [Guenther, 1997]. In addition, manual assignment of metadata and automated assignment of metadata are also two sides of a spectrum of data-recording approach. To assign metadata for each entity is a time-consuming, and human-intensive work. How to automate or semi-automate this work is a basis for developing effective digital libraries.

The main purpose of Information Extraction (IE) systems is to extract information from documents based on predefined templates. What kind of information is appropriate is a domain-dependent issue. For example, the information conveyed by business news or by terrorism news are very different. As a result, the predefined templates play an important role in IE systems. In fact, the predefined templates are the so-called metadata format. The joint efforts on IE and metadata will benefit both sides.

The sections what follows will describe what the metadata is and the metadata in usage. Section 3 will talk about information extraction and identify the relationship between metadata and information extraction. Section 4 will propose a parsing algorithm for initial analysis of documents, which is about to be further processed. Section 5 will carry out some experiments for the proposed parsing algorithm and discuss the experimental results. Section 6 is a brief conclusion.

2. Metadata

Metadata is data about data. That is to say, metadata is used to describe other information based on some rules or policies. In fact, various metadata formats are used in daily life. For example, each person has an ID card which is used to identify oneself and the ID card usually conveys the information such as ID number, name, birthday, birthplace, address, and parents. These data can be thought as metadata.

In order to make the readers or users convenient to find the books in libraries, each book has been cataloged in Machine-Readable Cataloging (MARC) [LOC, 1980] format based on Anglo-American Cataloging Rules, 2nd edition (AACR2) [1978]. Take the book "The Electronic Libraries" by Kenneth E. Dowlin as an example. Figure 1 shows its MARC format. MARC is very complicated and its different fields record the corresponding information for various retrieval purposes. From this viewpoint, the MARC can be regraded as a kind of metadata, too. However, MARC is for machine but not for human beings. It is necessary to map MARC to other metadata for human beings.

While the WWW becomes a fashion, the number of persons connecting to Internet increases very quickly. Many companies, organizations, universities, and persons publish their own homepages and archive a great deal of documents as online resources. In order to help people find appropriate information in the cyberspace, new services such as search engines and subject directories make their appearance. For example, Alta Vista (http://altavista.digital.com/) is a famous service provider and it now allows multilingual search. Yahoo (http://www.yahoo.com/) is a popular subject directory. It classifies registered homepages in numerous subjects. However, many shortcomings exist in the current working models of these service providers.

```
001     83021957 //r91
005     19911024125216.4
008     831004s1984     nyua       b      00110 eng   cam a
010     83021957 //r91
020     0918212758 (pbk.) :|c$24.95
040     DLC|cDLC|dDLC
050 00  Z678.9|b.D68 1984
082 00  025/.04|219
090     Z/678.9/D68/1984///1410222AL/1415924CL/1453410CL/1733896CF
091     TUL|bAL|bCL|bCL|bCF
095     TUL|dZ678.9|eD68|y1984|t095|bAL|c1410222
095     TUL|dZ678.9|eD68|y1984|t095|bCL|c1415924
095     TUL|dZ678.9|eD68|y1984|t095|bCF|c1733896
095     TUL|dZ678.9|eD68|y1984|t095|bCL|c1453410
099     TUL|d|e|y|f|t091|b|c|x|z
100 10  Dowlin, Kenneth E
245 14  The electronic library :|bthe promise and the process /
          |cKenneth E. Dowlin
260   0 New York, N.Y. :|bNeal-Schuman Publishers,|cc1984
300     xi, 199 p. :|bill. ;|c23 cm
440   0 Applications in information management and technology
          series
504     Includes bibliographical references and index
650   0 Libraries|xAutomation
650   0 Information technology
910     8'93 D#139            MCL
```

Figure 1. MARC Format

In order to make a high-quality subject directory, many human efforts are devoted to classifying homepages in Yahoo. Although Alta Vista employs an automatic indexing mechanism, the full-text search service is not high-quality enough as we expect. To get rid of these problems, researchers from computer science circle, library science circle, and network technology circle meet together at Dublin, Ohio in

1995 to discuss a metadata, which is suitable to describe Internet resources. Finally, a brand-new metadata called Dublin Core Set [Weibel *et al*., 1995] takes its first glance at the world. After a serial of discussions, Dublin Core under the Warwick Framework is proposed [Weibel *et al*., 1997]. The number of elements also increases from 13 to 15 to include image materials.

From the user's point of view, the functions of metadata are location, discovery, documentation, evaluation, and selection. Based on this line, Dublin Core Set defines fifteen core elements for description of WWW resources. These 15 elements fall into three categories shown in Table1. Researches of Dublin Core suggest that the next version of HTML integrate these core elements into tag set. As a result, newly produced homepages will be easy to be searched by users. However, it is needed to develop some automatic or semi-automatic procedures to “catalog” these existed homepages or other untagged documents without large human efforts. Researches of information extraction cast light on the resolution to these problems.

In fact, there doesn’t exist one universal metadata format for different document-like-objects (DLOs). The Dublin Core just one metadata format, and its core elements are as simple as possible. It is designed for generic networked DLOs. That is to say, it is necessary to devise some kind of metadata format for other digital repositories, if we want to provide more precise services. For example, the THA metadata format consists of 31 elements and some sub-elements. [Chen, 1998] However, the degree of automatic processing for such kind of metadata is lower than that for Dublin Core. The relationship of complexity and automaticity is shown in Figure 2.

Table 1. Elements of Dublin Core

Content	Intellectual Property	Instantiation
Title	Creator	Date
Subject	Publisher	Type
Description	Contributor	Format
Source	Rights	Identifier
Language		
Relation		
Coverage		

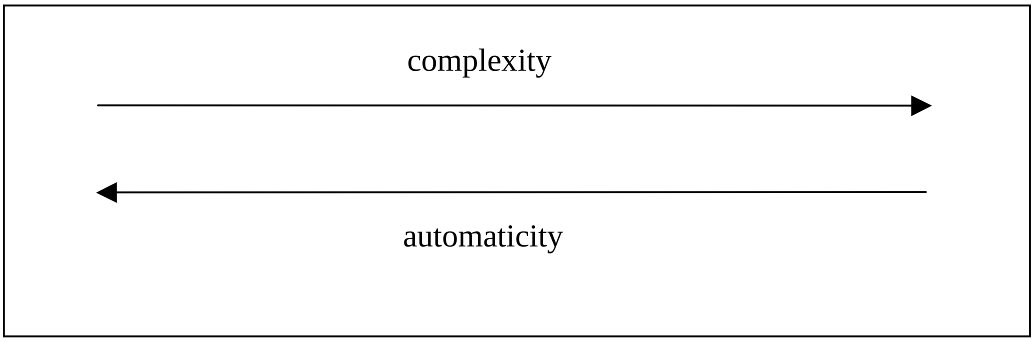

Figure 2. Complexity and Automaticity of Metadata Format

3. Information Extraction

Information Extraction (IE) is a task, which draws out some information from documents based on predefined templates. Researches of IE have to process documents much deeply than those of IR do. As Message Understanding Conference (MUC) describes, IE not only identifies the important constituents, but also the relationship among these constituents. Due to the specificity of task, extracting what kind of information is domain-dependent. For example, the target documents of MUC-5 are news articles about joint ventures and microelectronics and those of MUC-6 are news articles about management changes. [Appelt and Israel, 1997]

The tasks designed for MUC-6 are name identification, coreference resolution, and scenario template. The results of the first two tasks are the core parts of the third task and the output of the first task is critical for the second task to solve the coreference phenomena. The performances of these systems are shown in the Table 2. [MUC, 1997]

Table 2. Performance of Various Systems in MUC-6

Tasks	Recall		Precision	
	average	highest	average	highest
Name Identification	90%	96%	90%	97%
Coreference Resolution	66%	75%	76%	86%
Scenario Template	45%	47%	65%	70%

In fact, the third task of MUC-6 is what an IE system should carry out. A predefined template is just as a collection of attribute-value pairs. Thus, it is easy to know that the templates play the roles of metadata formats but with different faces. IE systems have to link the appropriate information drawn out from documents to the respective attribute.

A basic IE system for solving the aforementioned tasks consists of tokenization module, stemming module, lexical analysis module, and syntactic analysis module. For specific domain, the domain semantic knowledge is also a crucial component. Researches of Natural Language processing (NLP) have developed many high-performance analysis systems. The performance of tokenization module is about 98% correct rate [Palmer and Hearst, 1994]. The difficulty of this part is to distinguish whether periods are full-stop or part of abbreviations. The performance of stemming module is also good enough for IE task. The famous approach to stemming is two-level morphology [Koskenniemi, 1983]. As for lexical analysis module, this is the most improved part of researches of NLP in recent years. The various approaches to tagger are proposed. Some are probabilistic taggers [Church, 1988]; some are rule-based taggers [Brill, 1992]; some are hybrid taggers [Voutilainen, 1993]. The performance of taggers now approaches 98%. The syntactic analysis module is the most challenging work in joint researches of NLP, IR, and IE.

From the viewpoint of NLP, the correct and complete parse tree is very important. However, for applications like IR and IE, time is the most critical factor. How to leverage time and correctness factors in IR and IE systems is important research issue. Many researchers have proposed numerous syntactic analysis approaches. Fidditch [Hindle, 1983] is a deterministic parser based on hundreds of rules and some attachments of constituents are not resolved (this is why it is called a *partial* parser). Therefore, the output of Fidditch is usually a forest rather than a tree. Brill and Marcuss [1992] proposes an automatic procedure to acquire phrase structures from text corpora using distribution analysis and then use these phrase structures to parse unseen simple sentence with 4 to 15 words. The advantage of Brill and Marcus' approach is that they use tagged corpus only. However, they also regard these acquired phrase structures as grammar rules and manage to parse sentences using these rules. To get rid of the difficulties and to take executing time into account, a fast and new approach to parsing sentence is proposed and some experiments are conducted to show its performance in the following sections.

4. A Parsing Algorithm

The idea is when a sequence with the part-of-speech tags is seen, a human being, in general, has a capability to divide them into two parts and then repeatedly to divide each part into another two parts. For example, a sentence "stop electing life peers" with parts of speech sequence, {VB VBG NN NNS}, will be repeatedly divided as the Example 2 shows.

Example 1 The dividing process of sentence "stop electing life peers" is shown in Figure 3.

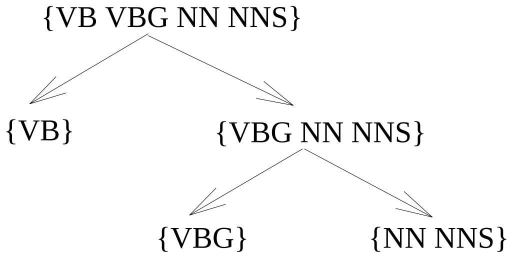

Figure 3. Repeatedly Dividing Sentence

As a result, the corresponding tree structure of Example 2 is [[VP : stop_VB [VP : electing_VBG [NP : life_NN peers_NNS]]] ._.], where the tag VB denotes base form of verb, VBG present participle, NN singular common noun, and NNS plural common noun [Johansson, 1986]. Such kind of procedure has many advantages. Firstly, the tree produced under this procedure is a binary-branch tree structure. It is easy to transform binary tree to other tree structure defined by grammar theory. From the point of view, the binary-branch tree can be seen as a canonical tree. Secondly, this procedure is language-dependent. That is to say, with tagged text corpora, it is easy to construct a syntactic analysis module. The human cost will be reduced to a great extent. Thirdly, since no deterministic or probabilistic grammar rules is needed in the proposed approach, the time cost to build a parser is greatly reduced.

For a given part-of-speech tags sequence, t_1, t_2, ..., t_{n-1}, t_n, if we want to separate these tags into two pieces, how do we choose the cutting point? Consider the *i*'th position between t_i and t_{i+1}. If this position is a plausible boundary, the number of possible tags following t_i and the number of possible tags followed by t_{i+1} will be much larger than other tag pair. Furthermore, the number of co-occurrance of (t_i, t_{i+1}) will be lower. Formally, we define forward entropy and backward entropy of a part of speech as the follows to fulfill the aforementioned concept.

Definition 1. Forward Entropy (*FE*). The forward entropy of a tag, t_i, is

$$FE(t_i) = -\sum_{t_j \in Tagset} P(t_j \mid t_i) \log P(t_j \mid t_i) .$$

Definition 2. Backward Entropy (*BE*). The backward entropy of a tag, t_i, is

$$BE(t_i) = -\sum_{t_j \in Tagset} P(t_i | t_j) \log P(t_i | t_j) .$$

In general, if t_i and t_{i+1} are really dominated by a nonterminal, the forward entropy of t_i and the backward entropy of t_{i+1} will be much lower. This is because what can follow t_i and what can be followed by t_{i+1} are more constrained.

Definition 3. Mutual Information (*MI*). The mutual information of two tags, t_i and t_j, is $MI(t_i, t_j) = \log \frac{P(t_j|t_i)}{P(t_j)} = \log \frac{P(t_j, t_i)}{P(t_i)P(t_j)}$.

If t_i and t_j usually appear together in corpus, the joint probability $P(t_i, t_j)$ will be larger than the $P(t_i) \times P(t_j)$. As a result, $MI(t_i, t_j)$ will be much larger than zero. On the contrary, if t_i and t_j distribute complementarily, $MI(t_i, t_j)$ will be much less than zero. According to Definitions 1, 2, and 3, we define the separability of a position between two tags, t_i and t_j, to be:

Definition 4. Separability Measure (*SM*).

$$SM(t_i, t_j) = FE(t_i) + BE(t_j) - MI(t_i, t_j).$$

Based on Definition 4, the higher the separability of a position is, the more plausible the position be to a boundary. Therefore, after calculating the separability, we repeatedly separate the sentence into two parts on the highest separability position in each iteration. Example 2 shows this case.

Example 2 The separability of positions in the sentence "Jack_NP Young_NP is_BEZ also_RB a_AT doubtful_JJ starter_NN next_AP year_NN" is shown in Figure 4 and the corresponding tree structure is shown in Figure 4.

[[[[Jack_NP Young_NP] [is_BEZ also_RB]] [[a_AT doubtful_JJ] starter_NN]] [next_AP year_NN]]

It is easy to train the value of *FE*, *BE*, and *MI* from tagged text corpora. For example, LOB Corpus [Johansson, 1986] contains one million words of British English with part-of-speech tags and Brown Corpus [Francis and Kucera, 1979] contains one million words of American English with part-of-speech tags. With these definitions in mind, we can cut a given sentence into pieces and then build the corresponding parsing tree.

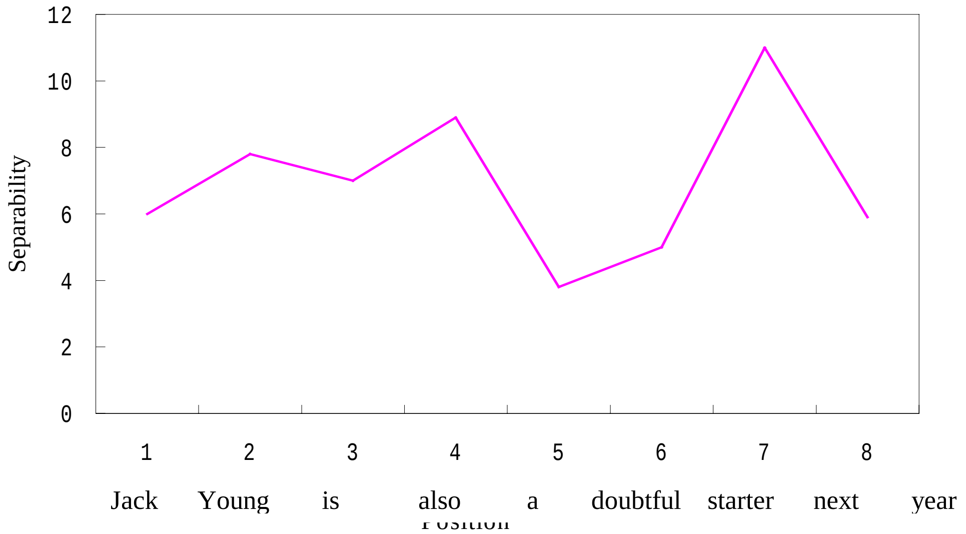

Figure 4. The Separability of a Sentence

5. Experiment and Evaluation

To evaluate the performance of a tool is not an easy task. In this section, we discuss two performance measures proposed in [Black *et al.*, 1991], PARSEVAL and Crossing-Bracket. Generally speaking, PARSEVAL is a strict measure. On the contrary, Crossing-Bracket is a loose measure.

Definition 5. PARSEVAL. The PARSEVAL measure consists of two criteria, which is popularly used in information retrieval.

$$\text{Precision} = \frac{\#\,\text{of correct constituents}}{\#\,\text{of produced constituents}}$$

$$\text{Recall} \;=\; \frac{\#\,\text{of correct constituents}}{\#\,\text{of constituents in treebank}}$$

For example, the evaluated parse tree and the correct parse tree of sentence "the daring ugly dog barked at Mary" are [[the_AT [daring_JJ ugly_JJ]] [dog_NN [barked_VBD [at_IN Mary_NP]]]] and [[the_AT [daring_JJ [ugly_JJ dog_NN]]] [barked_VBD [at_IN Mary_NP]]], respectively. The tag AT denotes determiner, JJ adjective, NN singular common noun, VBD past tense verb, and NP proper noun. Because the evaluated parse and the correct parse are binary trees, the numbers of constituents of them are the same as each other. Therefore, the recall and the precision will be equal in this experiment. In the above example, the number of constituents is 6. The number of correct constituents in evaluated parse is 2, i.e., the constituent [barked at Mary] and [at Mary]. Therefore, both the recall and the precision are 2/6 = 0.33.

Definition 6. Crossing-Bracket. Crossing-Bracket is the number of constituents which are not compatible with the standard parse.

Definition 7. Accuracy. Accuracy is the ratio of non-crossing bracketed constituents in a produced parse against all of the constituents.

Use the aforementioned example. The constituents in correct parse are [at Mary] [barked at Mary], [ugly dog], [daring ugly dog], [the daring ugly dog] , [the daring ugly dog barked at Mary]. The constituents in evaluated parse are [at Mary] [barked at Mary], [dog barked at Mary], [daring ugly], [the daring ugly], and [the daring ugly dog barked at Mary]. The number of *crossing-bracketed* constituents is 1, that is, [dog barked at Mary]. Figure 5 gives a clear overview of this idea.

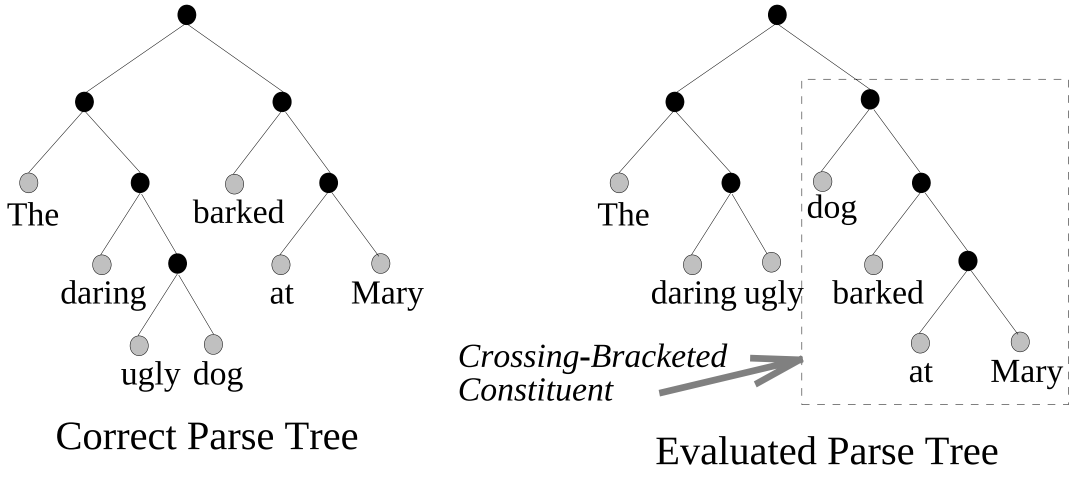

Figure 5. Crossing-Bracket Measure

In order to evaluate the performance of the proposed model, we use the SUSANNE Corpus [Sampson, 1995] as the testing corpus. SUSANNE Corpus is a tree bank of Brown Corpus [Francis and Kucera, 1979], but the former only contains one tenth of texts of the latter. The Corpus consists of four kinds of texts: 1) A: press reportage; 2) G: belles letters, biography, memoirs; 3) J: learned writing; and 4) N: adventure and Western fiction. The Categories of A, G, J, and N are named from each of the Brown Corpus. Each Category consists of 16 files and each file contains about 2000 words. Table 3 gives an overview of the Susanne Corpus. The details can refer to [Sampson, 1995].

Figure 5 illustrates the histogram of sentence length. Figure 6 and Figure 7 show the experimental results with different evaluation measures. Figure 6 demonstrates the performance using accuracy measure. Figure 7 the precision (recall) of the proposed parsing algorithm. The crossing bracket measure and PARSEVAL measure show the different behaviors. The crossing bracket measure is loose. The number of non-crossing bracketed constituents is much larger than that of crossing bracketed constituents. In contrast, the number of exactly matching constituents is much less than that of all constituents. Table 4 shows the performance of three different ranges 4-40, 4-25, and 10-20 of sentence length. In general, the crossing brackets per sentence in this part is about 2.5, the accuracy is about 0.81, and the precision is about

0.33. There is no significant difference among the three ranges. This demonstrates the proposed algorithm has the uniform power in processing SUSANNE Corpus within the range of numerous sentence length.

Table 3. The Overview of Susanne Corpus

Categories	Files	Paragraphs	Sentences	Words
A	16	767	1445	37180
G	16	280	1554	37583
J	16	197	1353	36554
N	16	723	2568	38736
Total	64	1967	6920	150053

Table 4. Parts of Experimental Results

Sentence Length Range	4 - 40	4 - 25	10 - 20
# of Sentences	5905	4373	2400
Average of Sent Length	19.09	14.71	15.02
Crossings Per Sentence	2.655	2.327	2.482
accuracy	0.821	0.801	0.804
Recall (Precision)	0.319	0.355	0.316

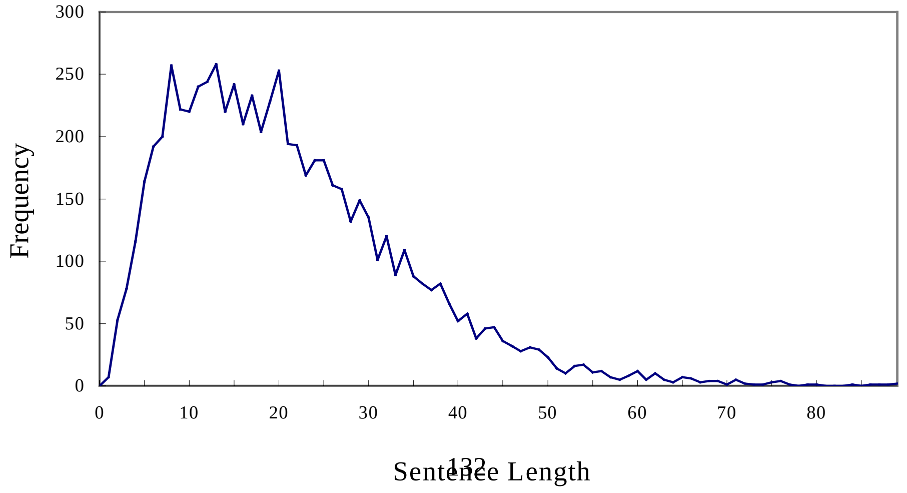

Figure 5. Histogram of Sentence Length

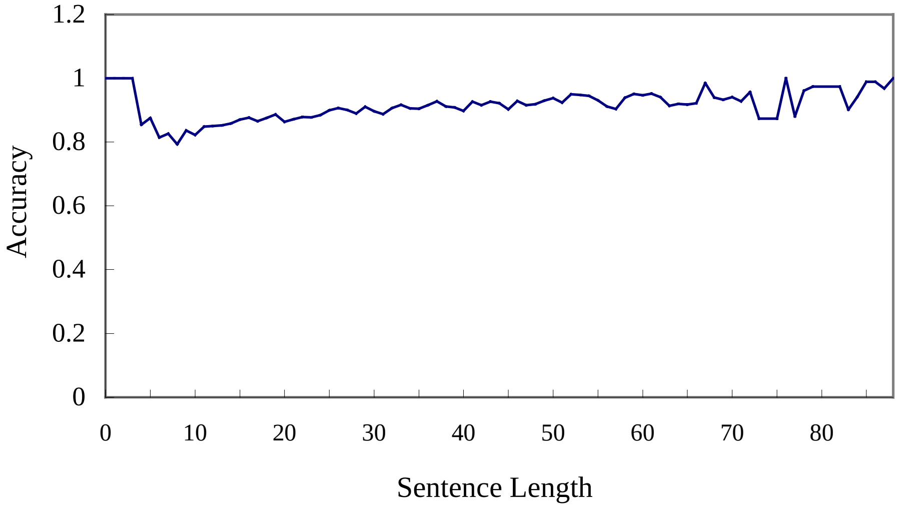

Figure 6. Accuracy w.r.t. Sentence Length

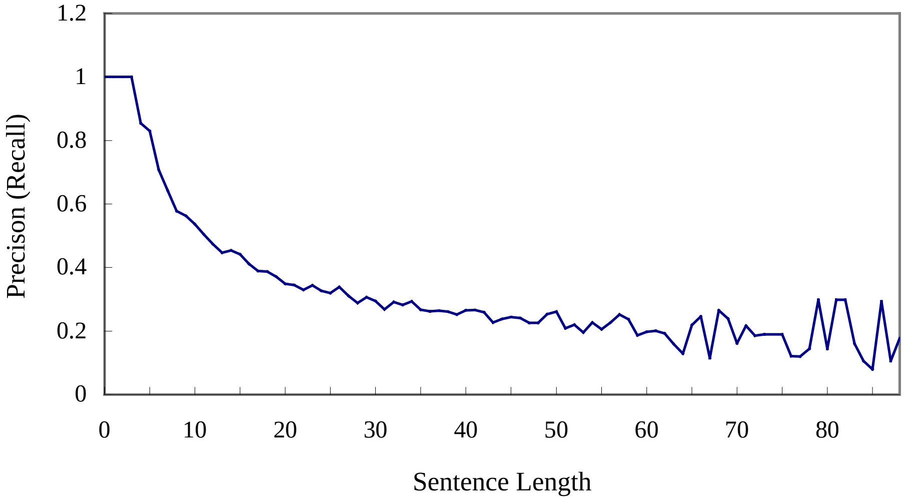

Figure 7. Precision (Recall) w.r.t. Sentence Length

6. Concluding Remarks

Both IR systems and IE systems help users to resolve the information need. Under an open environment like Internet, the quantity of heterogeneous documents increases quickly. Some sorts of metadata are critical to produce high-quality search results. Regarding digital libraries as an extension of IR systems, integration of researches on metadata and IE will effectively promote the service quality of digital libraries. This paper not only discusses the relations among metadata, IE, and IR, but also describe how to apply NLP technology to automating or semi-automating the IE tasks.

A parsing algorithm capable of fast analyzing documents is proposed in this paper. The main advantage is that only parts of speech information is needed in constructing a parsing module based on the proposed algorithm. Since the technique of parts of speech tagging is mature [Church, 1988; Cutting *et al*., 1992; Brill, 1992], a parsing module is easily constructed for any natural language provided a large scale tagged text corpora is available. Two different performance measures are used to evaluate the proposed parsing algorithm. Although the performance with respect to PARSEVAL measure is about 0.33, the performance with respect to accuracy measure is about 0.81.

References

AACR2 (1978). *Anglo-American Cataloging Rules*, second edition, American Library Association, Chicago, 1978.

Appelt, D.E. and Israel, D. (1997). *Tutorial on Building Information Extraction Systems*, 1997, Washington, DC, p. 4.

Black, E. et al. (1991). "A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars," Proceedings of the Workshop on Speech and Natural Language, 1991, pp. 306-311.

Brill, E. (1992). "A Simple Rule-Based Part of Speech Tagger," *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy, pp. 152-155.

Brill, E. and Marcus, M. (1992). "Automatically Acquiring Phrase Structure Using Distributional Analysis," *Proceedings of the DARPA Conference on Speech and Natural Language*, pp. 155-159.

Chen, Hsueh-hua (1998). "Metadata for Taiwan Historical Archives," Workshop of Digitization of Taiwan Historical Archives, Taipei, 1998. (in Chinese)

Church, K. (1988). "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas, pp. 136-143.

CIMI (1997). Consortium for the Computer Interchange of Museum Information, (URL: http://www.cimi.org/).

Cutting, D.; Kupiec, J.; Pedersen, J and Sibun, P. (1992). "A Practical Part-of-Speech Tagger," *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy, 1992, pp. 133-140.

Francis, N. and Kucera, H. (1979). *Manual of Information to Accompany a Standard Sample of Present-day Edited American English, for Use with Digital Computers*, Department of Linguistics, Brown University, Providence, R. I., U.S.A., original ed. 1964, revised 1971, revised and augmented 1979.

Guenther, R. (1997). "Dublin Core Qualifiers/Substructure," 1997, (URL: http://www.loc.gov/marc/dcqualif.htm).

Hindle, D. (1983). "User Manual for Fidditch, a Deterministic Parser," *Naval Research Laboratory technical Memorandum 7590-142*, Naval Research Laboratory, Washington, D.C.

Johansson, S. (1986). *The Tagged LOB Corpus: Users' Manual*, Bergen: Norwegian Computing Centre for the Humanities, 1986.

Koskenniemi, K. (1983). *Two Level Morphology*, Ph.D. Thesis, University of Helsinki, 1983.

LOC (1980). *MARC Formats for Bibliographic Data*, Library of Congress, Automated System Office, 1980.

MUC (1997). *Message Understanding Conference*, (URL:http://tipster.org/muc.htm).

NSF/ARPA/NASA (1992). Research on Digital Libraries: Joint Initiative of the National Science Foundation (NSF), the Advanced Research Projects Agency (ARPA), and the National Aeronautics and Space Administration (NASA), Request for Proposals, 1992, (URL: http://farside.gsfc.nasa.gov/ISTO/DLT/nsf_info.html).

Palmer, David D. and Hearst, Marti A. (1994). “Adaptive Sentence Boundary Disambiguation,” *Proceedings of Applied Natural Language Processing*, 1994, pp. 78-83.

Sampson, G. (1995). *English for the Computer*, Oxford University Press.

Voutilainen, A. (1993). “NPTool, a Detector of English Noun Phrases,” *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio, 1993, pp. 48-57.

Weibel, S.; Cathro, W. and Iannella, R. (1997). "The 4th Dublin Core Metadata Workshop Report," D-Lib Magazine, June 1997, (URL: http://www.dlib.org/dlib/june97/metadata/06weibel.html).

Weibel, S.; Godby, J. and Miller, E. (1995). “OCLC/NCSA Metadata Workshop Report,” 1995, (URL: http://www.oclc.org:5046/conferences/metadata/dublin_core_report.html).