Natural Language Processing and Digital Library Management System

Krishna Karoo

Department of Computer Science, Science College Pauni, India

Abstract: ***The study of Natural Language Processing (NLP) has been developed since last 50 years or so, producing the bunch of now mature technology such as automatic morphological analysis, word sense, disambiguation, parsing, anaphora resolution, natural language generation, named entity recognition, etc. The rapid increase of large digital collections and the emerging economic value of information demand efficient solutions for managing the information which is available, but which is not always easy to find. This paper presents the requirements for handling documents in digital libraries and explains how existing Natural Language Processing (NLP) technology can be used to build the task of document management.***

Keywords: Digital Libraries, Document Management, NLP applications for Digital Libraries, Metadata, Content Processing, Content analysis, Automatic classification, Named entity recognition for Digital Libraries, Thesauri, Document and information retrieval

1. Introduction

The study of Natural Language Processing (NLP) has been developed since last 50 years and so from the first machine translation and information retrieval applications to the present. These two areas of research have been far reaching and spreading widely. In the process of resolving issues of understanding natural language, for both translation and retrieval, many sub-areas of NLP have emerged: automatic morphological analysis, word sense disambiguation, parsing, anaphora resolution, natural language generation, named entity recognition, etc.

Now a day's research in Natural Language Processing (NLP), attention has been shifted from machine translation over to different versions of Information Retrieval (IR) applications. The increasing availability of large collections of digital documents has spurred interest in devising useful technology to handle these. Specifically, the notion of "digital libraries" has emerged, with specific architecture and functionality. This is an area where many mature Natural Language Processing (NLP) applications can be brought into play. It is an area mostly associated with Information Retrieval (IR), which has traditionally used little Natural Language Processing (NLP) and yet produced efficient tools; methods needed to include more sophisticated, Natural Language Processing (NLP)-based approaches were, up to recently, beyond the reach of IR systems. But digital libraries are much more than simply Information Retrieval (IR).

Objectives

1) Describe the issues relating to the task of managing a digital library
2) Explore various Natural Language Processing (NLP) applications which can be applied to the task
3) Identify new research problems related to these issues

2. Background

2.1 Digital Libraries

Digital collections existed long before the advent of the Web and the coinage of the term "digital library". NetLib (http://www.netlib.org/), created in 1985, contains a collection of freely available software, documents, and databases of interest to the numerical, scientific computing, and other communities. The Perseus project (http://www.perseus.tufts.edu/hopper/) was created in 1985 to host a collection of 2 resources on Ancient Greece: documents, images of artefacts, maps and the like, all linked together to allow a better understanding of Ancient Greek texts. Cornell University's e-prints archive (http://arxiv.org/), formerly the Los Alamos E-print Archive, dates from 1991.

An early definition, still cited today, comes from Borgman (2000, 42), in which a digital library is as follows:

> *... a set of electronic resources and associated technical capabilities for creating, searching, and using information. In this sense they are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium (text, images, sounds; static or dynamic images) and exist in distributed networks. The content of digital libraries includes data, metadata that describe various aspects of the data (e.g. representation, creator, owner, reproduction rights), and metadata that consist of links or relationships to other data or metadata, whether internal or external to the digital library.*

2.2 Document Management

The metaphor chosen to describe collections of digital content has been the library, not only because of the fact it houses a collection of documents, but also because its aim is that of the traditional library: to allow its users to access its contents (a set of digital resources) efficiently. It follows naturally that the desired functionality from a digital library can be inspired by its traditional counterpart. Document management as performed in a traditional library setting (as described in Lancaster, 2003, for example) involves a series of steps. First, from an initial potentially infinite source of resources (the Web, for example), a selection is made by the library's managers to retain a certain type or a certain number of resources, hereafter referred to as documents, to make up the library's collection.

On the representational axis, these documents need to be represented by a formal description, including title or name, author or creator, source, location, format, etc., i.e. with descriptive metadata. The descriptions are then inserted in a local organizational system: a catalogue; they may have additional metadata attached to them, such as index terms or classification codes, a short summary or description (semantic metadata). On the physical axis, the documents (or their representation) are stored (or accessible via hyperlinks). Finally, functionality is provided to the user for searching or accessing these documents: a search engine, a browse able index or classification scheme, etc., which provide access to the descriptions and/or he documents. In addition, the library, or rather its agents, can disseminate information (such as new acquisitions) to its users. The steps are thus: document selection and acquisition, description, classification, indexing and abstracting, storage, and distribution or presentation to users. In the digital realm, this so-called "document chain" is a closed one, as users are very often document creators themselves. In addition, with today's facilities for document annotation and tagging, the user may even provide descriptions of various kinds, thus taking an even more important role in the chain, which may not be best described as a chain at all.

2.3 Metadata

From a library and information science (LIS) perspective, metadata corresponds to cataloguing information; that is, the description of a resource by (mainly) its physical or "external" attributes: title, author, publication or creation date, format, length (page numbers for texts, minutes for video and audio), etc. From a computer science perspective, an early definition:

> *Metadata is data associated with objects which relieves their potential users of having to have full advance knowledge of their existence or characteristics. It supports a variety of operations. A user could be either a program or a person. (Dempsey & Heery, 1998)*

Until the middle of the 1990s, the term was used by the data management and systems design communities with a narrower interpretation, relative to a set of standards (Gilliland-Swetland, 2000). Today, its meaning extends to normalized descriptions of resources, digital or other (catalogues, indexes, archival search tools, museum documentation, etc.). The Dublin Core metadata scheme (http://dublincore.org/documents/dces/), although intended to describe any online resource, contains much of the basic information that librarians recognize as cataloguing information. Its fifteen elements are: contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title and type. The subject and description metadata element correspond to indexing and summarization. Supplying values for these elements requires more than a perusal of superficial document properties, but rather a relatively thorough examination and description of the resource's topic, focus and content.

2.4 NLP and Document Management

The role that NLP can play in document management was realized early on (e.g. Masterman et al., 1958; Sparck Jones, 1967), particularly for document retrieval. The interest is growing (see for example Ambroziak & Woods, 1998; Strzalkowski 1999; Voorhees 1999; Perez-Carballo & Strzalkowski 2000; Oard et al., 2001; Todirasçu & Rousselot, 2001; Ruch, 2003; Radev & Lapata, 2008; Kastner, 2009). There are important links to be made with the semantic Web, aimed at improving retrieval based on semantic grounds rather than on the presence of character strings in documents. See for instance the International Conference on Digital Libraries and the Semantic Web (http://www.icsd-conference.org/). A new development, with the advent of powerful players like Google and the like, is that there are very important stakes involved, due to the growing economic value of digital information. Practically all NLP applications are relevant and potentially useful in a digital library setting. In particular, methods for information retrieval are an integral part of search engines, and as such are incorporated in virtually any digital library along with all supporting technologies such as word-sense disambiguation, etc.

3. Overview of NLP Tools in Document Management

This section sketches the spectrum of NLP applications for document management, grouped according to four aspects:

1) Resource acquisition (including creation, representation and storage)
2) Content processing
3) Getting users in touch with documents
4) Knowledge organization tools

3.1 Resource acquisition

This aspect covers issues dealing with the acquisition of resources and the related questions of the representation of document files which are sensitive to language. A library's collection is never final; it is continually augmented by newly acquired material. Which material is added is determined by library policy, based on a number of criteria. Leaving economic matters aside, the criteria may include the following:

1) Topic (e.g. ornithology for a bird-watching club documentation center; business-related literature for a financial institution's library);
2) Genre (biographies or novels for a public library; conference proceedings for a university or research library; personal correspondence for an archival library; movies for a cinema school's library);
3) Intended audience (picture books for a preschool library; junior dictionaries for a school library);
4) Author (for government libraries).

Documents can be added to a digital collection by downloading, creation, digitization, transformation (from one format to another), etc.

3.1.1 Acquiring documents

In some cases, the acquisition of new documents to be added to a digital library can be automated using NLP tools. This is especially true when the selection criteria involve topic: a profile can be defined which expresses the selection criteria for the digital library, as features of the documents; new documents' contents can compared to the profile and processed by an automatic classification algorithm. Joorabchi & Mahdi (2008) describe an implementation of such functionality for a national repository for course syllabi (see also references therein). A very similar task is also performed by so-called « information-filtering systems » (see among others Belkin & Croft, 1992, Hanani et al., 2001), which intervene between an automated retrieval system and a user, to restrict the number of documents retrieved.

3.1.2 Determination of proper processing tools

Tools which will be used to process the documents, for example term extractors, part-of-speech taggers, summarizers, etc, are language-sensitive: German texts for instance require different tools than Chinese texts. It is a reasonable assumption in today's understanding of digital libraries that they are intended to be multilingual. To optimize the overall functioning of the library management system, it is desirable to include in the system functionalities for the automatic identification of language and encoding. Such systems have been developed in the past 15 years, based on character n-gram profiles. Řehůřek & Kolkus (2009) provide an up-to-date presentation in the context of the Web.

3.1.3 Document description

To represent and store documents in a digital library, it is necessary to produce some sort of record by which they are accessed. This corresponds to a traditional library's bibliographic entry, or a metadata record (i.e. descriptive metadata). This record is typically produced explicitly, either hand coded or automatically produced by extracting metadata from the resource. No semantics is involved and usually very little NLP technology. However, the normalization of author names and titles is a reasonable objective, and would require NLP tools similar to those for the normalisation of named entities (see for instance Andréani & Lebarbé, 2010). See also Kanhabua & Nørvåg (2008) on automatic means of determining a timestamp for documents which lack one. Also, one can imagine including here the results of automatic identification of document language and encoding, or of date formats. The descriptive or "physical" metadata described above is often not sufficient, or not ideal, for retrieval by a library's users. Additional metadata can be produced automatically by content processing.

3.2 Content processing

Content processing is a major part of the document management endeavour. It consists in producing enhanced metadata descriptions, in order to facilitate document retrieval by users, in addition to the retrieval capabilities provided by full-text searching. Resulting metadata is to be included in the digital library's knowledge organisation system. Content processing implies performing an analysis of the linguistic and/or conceptual contents of the text documents, and produces appropriate representations for these documents (such as indexing terms, summaries, classification codes, etc.). Content processing thus covers the traditional tasks of classifying, indexing and summarizing documents. Classifying implies grouping together documents on similar topics, and usually makes use of a classification scheme (such as the Dewey Decimal Classification or the Universal Decimal Classification, etc.); its analog in the digital world would be the hierarchical presentations of directories. Indexing (which may be interpreted differently by different communities) involves here the description of documents with a short list of terms or keywords representing the main topics discussed in the document. Summarization yields a shortened form of documents in a (usually) narrative style.

3.3 Getting users in touch with documents

This aspect deals with the raison d'être of libraries: access to documents by users, either by their own initiative (retrieval) or by the information system's ability to broadcast news out to a community of users.

3.3.1 Document/information retrieval

In a traditional library setting, actual document retrieval is often preceded by a "reference interview", where a librarian tries to ascertain the exact information needs of the user and thus to develop a successful search strategy which will include online search as well as searches in other sources. In a digital library world, this initial phase is non-existent. Users refine their search strategy themselves, gradually, as a reaction to the responses of the system and to what they discover about the contents of the collection. In addition, certain features of the digital library system have been designed to simulate the broadening or sophistication of the search that a librarian would perform. And thus document retrieval in a digital setting is reducible to so-called "information retrieval". This is probably the best-researched field in document management. The presentation here will only aim to underline the array of NLP technology used (this is also addressed by Mustafa el Hadi, 2004).

3.3.2 Broadcasting documents to users

It is customary for an information service such as a library to issue bulletins to its users, informing them of new material or special events, when appropriate. This can be done through mailing lists, billboards, etc. The equivalent in the digital world is straightforward. What is novel here, however, is that bulletins can be tailored to individual user profiles. Specifically, new documents can be analysed (indexed, classified or summarised) and compared to a user profile consisting of user-supplied or system-supplied keywords; in the event of a match, users can be notified of these new documents through appropriate messaging technology (e-mail, RSS feed, etc.). Such a system is described in Morales del Castillo et al. (2009) while Gu et al. (2008) present a similar functionality to support learning.

3.3.3 Answering users' questions

A major part of every librarian's day involves answering questions for users. Some modern versions of such a reference service employ chat rooms and the like ("Ask-a-librarian" services), with a human librarian accessible over the internet. An even more modern take on the idea is to use a question answering system, such as in Mittal et al. (2005) or Bloehdorn et al. (2007). The task of relating users to documents is obviously at the core of a library's mission and of digital libraries' functionalities. NLP tools can assist in various ways, as has been illustrated so far. We now turn to an aspect which transcends document management tasks.

3.4 Knowledge organization tools

We refer here to linguistic resources used in the text management and processing tasks described above. The one that is most specific to document management is the thesaurus (other knowledge organisation tools relevant for digital libraries are presented in Soergel, 2009).

3.4.1 Properties of thesauri

Note that the term "thesaurus" means slightly different things to information professionals (librarians) and computer scientists, or to language educators for that matter. Loosely speaking, a thesaurus is some kind of synonym dictionary; in reality it is much more. It encodes not only synonymous terms but also hierarchical relationships (i.e. which terms are broader and narrower than a given term) and other types of semantic relationships, depending on the resource. Specifically, the "thesaurus" most used in NLP applications, WordNet, is not a thesaurus by LIS standards.

The LIS version of the thesaurus (defined by international standards ISO 2788 and ISO 5064) adopts a stricter definition of thesaural relationships.

These are restricted to only three types:

1) Hierarchy (broader/narrower terms or generic/specific terms, otherwise known as hypernym/hyponym terms)
2) Synonymy (semantic equivalents which may include spelling variants, shortened forms, etc.)
3) The so-called associative relationship, relating terms that are neither synonyms nor in a hypernym/hyponym relation, yet are related semantically.

Thesaural relations exclude (almost all) partitive (part-whole) relationships and others which are routinely introduced in ontologies.

3.4.2 Uses of thesauri in digital libraries

The content management tasks (automatic indexing, classification and summarization) can greatly benefit from knowledge sources such as thesauri, which encode semantic relationships among words and terms. The two most basic of these are the synonymy relation and the hypernym/hyponym relation. The two can be used to improve on content processing, such as indexing with more general or more specific terms, and bringing together synonymous expressions to enhance indexing or to allow generalizations in summarizing.

3.4.3 Automatic construction of thesauri

Attempts have been made to create thesauri by automatic means, to overcome the problem of the scarcity of appropriate resources. General language thesauri (such as WordNet and the like) offer a wide coverage, but have serious limitations in specialized domains. Specialized thesauri have the opposite flaw (often too narrow in scope), and are in addition fairly rare, often not available for a given specialized domain. To circumvent these problems, the automatic construction of a thesaurus is an endeavour that has been attempted by several researchers (see for instance Auger & Barrière, 2008 and others). The linguistic challenge lies in the automatic identification of semantic relations of synonymy, hypernymy/hyponymy, and other "essential" semantic relationships which may be difficult to characterize exhaustively. All of these present serious challenges. This research area is close to that of ontology learning and population from text.

4. A Closer Look at Some Challenges for Digital Library Management

The previous wide-ranging exposé has identified numerous possibilities for NLP applications in the context of digital library management. The rest of this chapter focuses on certain specific challenges met by digital libraries.

4.1 Named entity recognition and resolution

It is useful and often necessary to be able to determine when two similar variants of a named entity in fact designate the same one: John Smith, J. Smith, Pres. Smith, John Smith Jr., etc. Organization names can also vary: Acme Deliveries vs. Acme Deliveries Inc; IBM vs. International Business Machines; The John Hopkins University vs. John Hopkins; etc. This problem is compounded when names come from a foreign country, possibly through transliteration from a foreign language. This has long been recognized in library cataloguing and is the focus of sections in the Anglo-American Cataloguing Rules handbook (Joint Steering Committee for Revision of AACR, 2002). In the domain of scholarly publications, names of institutions, universities, research laboratories, etc. can manifest different variants. This presents a problem when one wants to identify named entities emanating from different sources: different publications, different libraries, in bibliographies from different documents, sometimes dictated by bibliographic styles. It is a problem for a number of endeavours and is indeed a topic of many research papers related to digital libraries.

4.2 Tools to assist OCR

Some challenges arise due to the digitization process of certain types of documents: namely, historical documents and so-called retrospective collections of modern digital media. Access to these is hampered by the poor quality of the OCR text. Tahmasebi et al. (2010) investigate the effects of OCR errors on word sense discrimination results on historical documents; evaluations are performed on The Times newspaper archive, with documents dating from 1785 to 1985. Allen et al. (2010) tackle the task of identifying sections and regular features of historical newspapers in order to improve the automatic classification of articles; the ultimate goal is to provide improved search services for these documents.

4.3 Search and retrieval

Improved search strategies are needed. Methods which favour precision (eliminating irrelevant items) are especially sought, as we see the development of topical digital libraries – where distinctions between documents can be finer-grained than on the Web as a whole (Bethard, 2009). On the other hand, to enhance recall, the integration of lexical resources such as thesauri and ontologies should be useful.

4.4 Retrieval of non-textual documents

One interesting aspect of digital libraries is that they bring together three formerly quite distinct disciplines, i.e. libraries, archives and museums. Digital resources in digital libraries are not limited to textual documents, nor to digital objects, but can include images, video, sound, and digital renderings of three-dimensional physical objects. The extraction of information from the text surrounding images can support automatic indexing of these images (see for instance Haruechaiyasak & Damrongrat, 2010), and the same can be applied to video, audio or multimedia resources (Da Sylva & Turner, 2005).

4.5 Genre-based processing

Genre-based processing (i.e. that which takes into account the genre or type of a document and can adjust accordingly) is an important issue that can be tackled by NLP means. For example, in automatic summarization, Saggion & Lapalme (2000) take advantage of the predictable structure of scientific articles to focus on certain sections from which to extract sentences which will appear in the final extract. Chieze et al. (2010) take a similar approach to handle specific types of legal documents (court judgement renderings, and intellectual property and tax law texts). The latter are examples of single-genre processing. To allow for processing of more than one genre would improve on existing, "off-the-shelf" technology which is geared towards a single genre.

5. Conclusion

The digital library setting represents an interesting opportunity for computational linguistics: it can use many new applications with great potential (notably, a great financial or economic potential, given the new economic value of information). Current focus on very large digital libraries may test the robustness of seemingly mature NLP technology.

In the past, syntax has played a large role in NLP development, notably in symbolic approaches to machine translation, where systems were developed with translation rules from one language's syntactic constructions to another. So far, syntax has played a very small part in NLP for document management (see however Spagnola & Lagoze, 2011). Research must now focus on computational semantics: lexical, phrasal and sentential semantics, and in even higher level units. Indeed, text linguistics or discourse analysis will drive new research, especially for summarization and certain approaches to classification. In the long term, the ultimate challenge will be to model more than merely the linguistic dimensions of digital library management, adding also cognitive, communicational, pragmatic, social or semiotic dimensions, etc. These can appeal to cognitive science and artificial intelligence in general; but even in the linguistic dimensions, challenges abound.

References

[1] Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., & Tsakonas, G. (2009). In M. Agosti, J. Borbinha, & S. Kapidakis (Eds.), Research and Advanced Technology for Digital Libraries. Proceedings of the 13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. Lecture Notes in Computer Science Volume 5714. Berlin; Heidelberg : Springer-Verlag.

[2] Andrews, J., & Law, D.G. (Eds) (2004). Digital libraries: policy, planning, and practice. Aldershot, Hants: Ashgate.

[3] Bontcheva, K., Maynard, D., Cunningham, H., & Saggion, H. (2002). Using Human Language Technology for automatic annotation and indexing of digital library content. In Proceedings of ECDL 2002 : European conference on research and advanced technology for digital libraries, Rome , 2002, vol. 2458 (pp. 613- 625). Berlin; Heidelberg : Springer-Verlag.

[4] Chengzhi, Z., & Dan,W. (2008). Concept Extraction and Clustering for Topic Digital Library Construction. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (pp. 299-302).

[5] Ferro, N. (2009). Annotation Search: The FAST Way. In M. Agosti, J. Borbinha, & S. Kapidakis (Eds.), Research and Advanced Technology for Digital Libraries. Proceedings of the 13th European Conference (pp. 15-26), ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. Lecture Notes in Computer Science Volume 5714. Berlin; Heidelberg : Springer-Verlag.

[6] Golub, K. (2006). Using Controlled Vocabularies in Automated Subject Classification of Textual Web Pages, in the Context of Browsing. TCDL Bulletin, 2(2). Retrieved October 8, 2010 from http://www.ieeetcdl.org/Bulletin/v2n2/golub/golub.html .

[7] Tuominen, K., Talja, S., & Savolainen, R. (2003). Multiperspective digital libraries: The implications of constructionism for the development of digital libraries. Journal of the American Society for Information Science and Technology, 54, 561-569.

[8] Witten, I. H., Bainbridge, D., & Boddie, S. J. (2001) Greenstone: open-source digital library software with end-user collection building, Online Information Review, 25(5):288-298.