The importance of context for Digital Libraries

Keynote Speech of the Lecture "Digital Preservation" in Máster of Information Management in Organization (1), University of Murcia, 18th April 2012

La importancia del contexto en las Bibliotecas Digitales.

Sándor Darányi
Swedish School of Library and Information Science. University of Borås
sandor.daranyi@hb.se

Resumen

El concepto de contexto es de gran importancia en la preservación digital. Este trabajo analiza su significado desde el punto de vista del acceso a los objetos digitales, combinando aspectos relacionados con la lingüística, la desambigüación terminológica en la recuperación de información y la categorización de textos. En estas áreas el contexto es clave para una correcta desambiguación y obtener mejores resultados. Por tanto, la preservación y posterior acceso de los objetos digitales debe contemplar la inclusión de la información adecuada sobre el contexto terminológico y social en el que se han generado dichos objetos.

Palabras clave: Preservación digital, Contexto de la Información, Recuperación y acceso a la Información

Abstract

The concept of "context" has great importance in digital preservation. This paper analyzes the meaning of context from the point of view of access to digital objects, combining linguistics, terminological disambiguation in information retrieval and text categorization aspects. In these areas, the context is a key element for successful disambiguation and get better results. Therefore, the preservation and subsequent access of digital objects should also consider the preservation of appropriate information about the terminology and social context in which these objects were generated.

Keywords: Digital preservation, Information in context, Information retrieval and access

1. Introduction

Today, context is one of those important words whose meaning, due to overuse, has become almost elusive. Good examples for this are the Blue Ribbon Task Force final report (Lavoie et al. 2010) and Skinner and Schultz (2010). For instance in his report on ontology building, Lennat (1998) mentions context 660 times. The key role this concept plays in daily communication is disambiguation and terminology bridge building between subject fields (Peterson et al. 2009), ultimately leading to vocabulary control in digital libraries.

Recent interest in digital preservation has steered context even more in the limelight. To exemplify its importance, I shall bring examples from linguistics, information retrieval and information extraction, and text categorization. What connects them is that together they are combined into advanced access to digital objects as it relates to language (terminology) change (Baker 2008). Also, context is an important component for any formal theory of digital libraries and digital preservation (Goncalves 2004, Dallas 2007, Flouris and Meghini …), as trustworthy repositories and the very concept of trustworthiness goes back to it.

2. The role of context in linguistics.

In linguistics, the study of meaning is called semantics, with word semantics constructing theories of word meaning, whereas sentence semantics is concerned with sentence meaning. Of the many theories of word meaning, Wittgenstein's version is quite famous: he was of the opinion that to learn a word's meaning one should observe its use (Wittgenstein 1953: 43). Namely, habitual usage provides indirect contextual interpretation of any term, an idea which independently returned first in Harris´ distributional hypothesis (Harris 1954), and then by Firth's famous maxim, "You shall know a word by the company it keeps" (1957:11).

One of the early applications of this observation in a modern automated text processing environment was Schütze's model which used index term co-occurrence as context to disambiguate homonyms by clustering (1998). Word senses were interpreted as groups (or clusters) of similar contexts of the ambiguous word. Words, contexts, and senses were represented in a so-called word space, a high-dimensional, real-valued space in which closeness corresponded to semantic similarity. Similarity in word space was based on second-order co-occurrence: two tokens (or contexts) of the ambiguous word were

assigned to the same sense cluster if the words they co-occurred with in turn occurred with similar words in a training corpus.

The general idea behind word space models is to use distributional statistics to generate high-dimensional vector spaces, in which words are represented by context vectors whose relative directions are assumed to indicate semantic similarity. This assumption is motivated by Harris' distributional hypothesis, which states that words with similar meanings tend to occur in similar contexts. According to this hypothesis, if we observe two words that constantly occur with the same contexts, we are justified in assuming that they mean similar things (Sahlgren 2005). A convincing model based on context is random indexing (Kanerva 1988), where the similarities in word use in a text corpus, such as a database, become apparent by such context vectors from a large-scale word-word co-occurrence matrix.

As Widdows states, context is vital for deciding which of the possible senses of a word is being used in a particular situation, a task known as disambiguation. Motivated by a survey of disambiguation techniques in natural language processing, he presents a mathematical model describing the relationship between words, meanings and contexts, giving examples of how context-groups can be used to distinguish different senses of ambiguous words (2003).

3. The role of context in advanced access to digital objects.

In language technology, and wherever it is used for text processing, context is the key to word sense disambiguation. The elimination of ambiguity, prominently by homonymy, and its impact on indexing, contributes to the quality of automatic document categorization and document retrieval. Beyond this, as indicated above, it is by virtue of keywords in context as encoded by their co-occurrence patterns into matrices that document categories and answers to user queries make sense.

We briefly mentioned vocabulary control as an important application context in the introduction. Vocabulary control is "a term in applied linguistics for the organization of words into groups and levels, especially as the outcome of frequency counts" (McArthur 1998).1 Problems of what qualifies as a word and should be counted, though, include, among others, orthographic ones (colour vs. color), homonymy (bear [animal] vs. bear [carry]), homographic examples (wind (air on the move) vs. wind (to turn, twist), and polysemy (mouth [body part] vs. mouth [of a river]). In a digital library, controlled vocabularies provide a way to organize knowledge for subsequent retrieval. They are used in subject indexing schemes, subject headings, thesauri and taxonomies. Controlled vocabulary schemes mandate the use of predefined, authorized terms that have been preselected by the designer of the vocabulary, in contrast to natural language vocabularies, where there is no restriction on the vocabulary.2

Consider some examples for terminology change over time e.g. from the BBC cataloguing team1: gramophone - record player; computer screen – monitor; duplicating machine – photocopier; European Coal and Steel Community – Common Market – European Economic Community or EEC – European Community or EC – European Union or EU; bicycle – bike – BMX. These are just a few of many more which make it an issue that in order to index and retrieve documents from certain periods, the right language must be used. In turn, usage (context) is vital for their recognition and handling.

Another typical task is to improve the accuracy of information retrieval by correctly identifying the context of the keyword being searched. E.g. substantial medical data, such as discharge summaries and operative reports are stored in electronic textual form. Databases containing free-text clinical narratives reports often need to be retrieved to find relevant information for clinical and research purposes. The context of negation, a negative finding, is of special importance, since many of the most frequently described findings are such. When searching free-text narratives for patients with a certain medical condition, if negation is not taken into account, many of the documents retrieved will be irrelevant. Hence, negation is a major source of poor precision in medical information retrieval systems. Previous research has shown that negated findings may be difficult to identify if the words implying negations (negation signals) are more than a few words away from them. In such cases, machine learning of content patterns for automatic identification of negative context in clinical narratives reports is useful, apart from speeding up manual knowledge engineering, context identification and information extraction tasks (Rokach et al. 2008).

Finally, a very interesting context-dependent application of text categorization for real-life scenarios is reported by (Sakaki et al. 2010), using Twitter for earthquake alerts. They suggest that an important characteristic of microblogging is its real-time nature. For example, when an earthquake occurs, people make many Twitter posts (tweets) related to it, which enables detec-

tion of earthquake occurrence promptly, simply by observing the tweets. To detect a target event, they devised a classifier of tweets based on features such as the keywords in a tweet, the number of words, and their context. Subsequently, a probabilistic spatiotemporal model locates the center and the trajectory of the event, which forms the basis of an earthquake reporting system in Japan. Because of the numerous earthquakes and the large number of Twitter users throughout the country, they can detect an earthquake by monitoring tweets with high probability (96% or more above scale 3). The system detects earthquakes promptly and sends e-mails to registered users; in fact notification is delivered much faster than the announcements that are broadcast by the Japan Meteorological Association.

4. Conclusions

For 21th century civilization, the concept of the embedding environment also known as context, and of contextual dependencies, is one of the most important. It is crucial not only for the communication and understanding of values but also for the digital preservation of complete technological, process and linguistic environments. Because the challenge is to move digital data into the future not just intact but in context while retaining its significance as research material, this makes the task a social, not just a technological problem (Thompson 2012).

Notes

(1) Editor's note: This job is an invited paper. It is adapted from the lecture "Digital preservation" in the Master of Information Management in Organizations of University of Murcia. The extensive professional, teaching and research experience of Prof. Sándor Darányi make him one of the foremost experts in Digital Preservation. He acted as the local coordinator of the EU project SHAMAN between 2008-2011. http://www.adm.hb.se/~sda/

References

Baker, A. (2008). Computational approaches to the study of language change. *Language and LInguistics Compass* 2, 2, 289-307.

Brocks, H., Kranstedt, A., Jäschke, G., Hemmje, M. (2010) *Modeling Context for Digital Preservation.* Heidelberg: Springer. 260 p.

Cruse, A. (2004). *Meaning in language: An introduction to semantics and pragmatics.* Oxford University Press: Oxford.

Dallas, C. (2007) An agency-oriented approach to digital curation theory and practice. In *International Cultural Heritage Informatics Meeting (ICHIM07): Proceedings*, J. Trant and D. Bearman (eds). Toronto: Archives & Museum Informatics. Published September 30, 2007 at http://www.archimuse.com/ichim07/papers/dallas/dallas.html [20-06-2012]

Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. Oxford University Press: London.

Giorgios, F. and Meghini, C. (2007). *Some Preliminary Ideas Towards a Theory of Digital Preservation*. CNR. http://puma.isti.cnr.it/dfdownloadnew.php?ident=/cnr.isti/2007-A2-152&langver=en&scelta=New-Metadata [10-11-2012]

Harris, Z. (1954). Distributional structure. *Word* 10, 23, 146–162.

Goncalves, M. A. (2004). *Streams, Structures, Spaces, Scenarios, and Societies (5S): A Formal Digital Library Framework and Its Applications*. PhD dissertation. Virginia Polytechnic Institute and State University: Blacksburg, Virginia.

Kanerva, P. (1988). *Sparse distributed memory*. The MIT Press

Lavoie, B., Berman, F., Smith Rumsey, A. (2010). Sustainable economics for a digital planet: Ensuring long-term access to digital information. Final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf [20-06-2012]

Lennat, D. (1998). *The dimensions of context-space*. CYCORP: Austin, TX.

Peterson, M., Zasman, G., Porter, J., Mojica, P., St. Pierre, E., Rogers, B. (2009). *Building a Terminology Bridge for Digital Information Retention and Preservation Practices*. SNIA DNF.

Rokach, L., Romano, R., Maimon, O. (2008). Negation recognition in medical narrative reports. *Information Retrieval*, 11, 499–538.

Sahlgren, M. (2005). An Introduction to Random Indexing. http://www.idi.ntnu.no/~gamback/teaching/TDT4138/sahlgren05.pdf [20-06-12]

Sakaki, T., Okazaki, M., Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. *Proceedings of the 19th International Conference on World Wide Web* (WWW-10), p.851-860.

Schütze, H. (1998). Automatic Word Sense Discrimination. Computational Linguistics, Volume 24,1 p.97-123

Skinner, K., Schultz, M. (2010). *A Guide to Distributed Digital Preservation*. Educopia Institute: Atlanta, GA.

Thompson, D. (2012). Things I wish I'd been told before I started…" DPC Digital Preservation Event, University College London, January 24, 2012. http://www..dpconline.org/events/details/38%E2%80%91studentconference?xref=38 [20-06-2012]

Widdows, D. (2003). A Mathematical Model for Context and Word-Meaning. *Proceedings of the Fourth International and Interdisciplinary Conference on Modeling and Using Context.* Stanford, California, June 2003, p.369–382.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell.