

KNOWLEDGE ORGANIZATION

KO

Official Quarterly Journal of the International Society for Knowledge Organization ISSN 0943 – 7444

International Journal devoted to Concept Theory, Classification, Indexing and Knowledge Representation

Contents

Call for papers for The International Society for Knowledge Organization (ISKO) Eighth (8th) International ISKO Conference (ISKO 8): Knowledge Organization and the Global Information Society. Hosted by the School of Library, Archive and Information Studies at University College London, Gower Street, London WC1E 6BT. July 13-16, 2004. 123

Guest Editorial

Guest Editor: Widad Mustafa El Hadi, University of Lille 3. A Special Issue of Knowledge Organization on Evaluation of HLT. Evaluating Human Language Technology: General Applications to Information Access and Management. 124

Articles

Ferret, Olivier; Grau, Brigitte; Hurault-Plantet, Martine; Illouz, Gabriel; Jacquemin, Christian; Monceaux, Laura; Robba, Isabelle; and Vilnat, Anne. How NLP Can Improve Question Answering. 135

Chen, Kuang-hua. Evaluating Chinese Text Retrieval with Multilingual Queries. 156

Sidhom, Sahbi and Hassoun, Mohamed. Morpho-syntactic Parsing for a Text Mining Environment: An NP Recognition Model for Knowledge Visualization and Information Retrieval. 171

Ibekwe-SanJuan, Fidelia and SanJuan, Eric. From Term Variants to Research Topics. 181

Bowker, Lynne. Information Retrieval in Translation Memory Systems: Assessment of Current Limitations and Possibilities for Future Development. 198

L'Homme, Denis; L'Homme, Marie-Claude; and Lemay, Chantal. Benchmarking the Performance of Two-Part-of-Speech (POS) Taggers for Terminological Purposes. 204

Fugmann, Robert. The Complementarity of Natural and Index Language in the Field of Information Supply. 217

KO Reports

Schneider, Jesper W. Emerging Frameworks and Methods: The Fourth International Conference on Conceptions of Library and Information Science (CoLIS4). The Information School, University of Washington. Seattle, Washington, USA. July 21-25, 2002. 231

Tennis, Joseph T. 6th Annual Open Forum on Metadata Registries. Santa Fe, New Mexico, USA. January 20-24, 2003. 234

Book Reviews

BADE, David. The creation and persistence of misinformation in shared library catalogs: Language and subject knowledge in a technological era. Urbana-Champaign, IL: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 2002. 33 p. ISBN 0-87845-120-X. (Occasional paper; no. 211). 236

HUNTER, Eric J. Classification made simple. 2nd ed. Aldershot, Haunts (UK): Ashgate, 2002. xi, 147p. ISBN 0-7546-0795-X (pbk). 237

HUYSMAN, Marleen and DE WIT, Dirk. Knowledge sharing in practice. Dordrecht: Kluwer Academic Publishers, 2002. 189 p. ISBN 1-4020-0584-9. 238

SAWONIAK, Henry, with the collaboration of Maria WITT. International Bibliography of Bibliographies in Library and Information Science and related fields. Volume II. 1979 – 1990. München: Saur, 1999. 3 Vols. (liii, 1208 p.). ISBN 3-598-11145-2. 240

KAO, Mary Liu. Cataloging and classification for library technicians. 2nd edition. New York : Haworth Press, 2001. xiv, 146 p. ISBN 0-7690-1063-1 (pbk). 241

Calls for Papers and Conferences 243

Contents page

O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, C. Jacquemin, L. Monceaux, I. Robba, A. Vilnat. (2002). **How NLP can improve Question Answering**. *Knowledge Organization*, 29(3/4). 135-155. 28 refs.

ABSTRACT: Answering open-domain factual questions requires Natural Language processing for refining document selection and answer identification. With our system QALC, we have participated in the Question Answering track of the TREC8, TREC9 and TREC10 evaluations. QALC performs an analysis of documents relying on multi-word term searches and their linguistic variation both to minimize the number of documents selected and to provide additional clues when comparing question and sentence representations. This comparison process also makes use of the results of a syntactic parsing of the questions and Named Entity recognition functionalities. Answer extraction relies on the application of syntactic patterns chosen according to the kind of information that is sought, and categorized depending on the syntactic form of the question. These patterns allow QALC to handle nicely linguistic variations at the answer level.

K.-H. Chen (2002). **Evaluating Chinese Text Retrieval with Multilingual Queries**. *Knowledge Organization*, 29(3/4). 156-170. 28 refs.

ABSTRACT: This paper reports the design of a Chinese test collection with multilingual queries and the application of this test collection to evaluate information retrieval systems. The effective indexing units, IR models, translation techniques, and query expansion for Chinese text retrieval are identified. The collaboration of East Asian countries for construction of test collections for cross-language multilingual text retrieval is also discussed in this paper. As well, a tool is designed to help assessors judge relevance and gather the events of relevance judgment. The log file created by this tool will be used to analyze the behaviors of assessors in the future.

S. Siddhom, M. Hassoun (2002). **Morpho-syntactic Parsing for a Text Mining Environment: An NP Recognition Model for Knowledge Visualization and Information Retrieval**. *Knowledge Organization*, 29(3/4). 171-180. 11 refs.

ABSTRACT: Sidhom and Hassoun discuss the crucial role of NLP tools in Knowledge Extraction and Management as well as in the design of Information Retrieval Systems. The authors focus more specifically on the morpho-syntactic issues by describing their morpho-syntactic analysis platform, which has been implemented to cover the automatic indexing and information retrieval topics. To this end they implemented the Cascaded “Augmented Transition Network (ATN)”. They used this formalism in order to analyse French text descriptions of Multimedia documents. An implementation of an ATN parsing automaton is briefly described. The Platform in its logical operation is considered as an investigative tool towards the knowledge organization (based on an NP recognition model) and management of multiform e-documents (text, multimedia, audio, image) using their text descriptions.

F. Ibekwe-SanJuan, E. SanJuan (2002). **From Term Variants to Research Topics**. *Knowledge Organization*, 29(3/4). 181-197. 21 refs.

ABSTRACT: In a scientific and technological watch (STW) task, an expert user needs to survey the evolution of research topics in his area of specialisation in order to detect interesting changes. The majority of methods proposing evaluation metrics (bibliometrics and scientometrics studies) for STW rely solely on statistical data analysis methods (co-citation analysis, co-word analysis). Such methods usually work on structured databases where the units of analysis (words, keywords) are already attributed to documents by human indexers. The advent of huge amounts of unstructured textual data has rendered necessary the integration of natural language processing (NLP) techniques to first extract meaningful units from texts. We propose a method for STW which is NLP-oriented. The method not

KNOWLEDGE ORGANIZATION

KO

Official Quarterly Journal of the International Society for Knowledge Organization ISSN 0943 – 7444

International Journal devoted to Concept Theory, Classification, Indexing and Knowledge Representation

only analyses texts linguistically in order to extract terms from them, but also uses linguistic relations (syntactic variations) as the basis for clustering. Terms and variation relations are formalised as weighted di-graphs which the clustering algorithm, CPCL (Classification by Preferential Clustered Link) will seek to reduce in order to produce classes. These classes ideally represent the research topics present in the corpus. The results of the classification are subjected to validation by an expert in STW.

L. Bowker (2002). **Information Retrieval in Translation Memory Systems: Assessment of Current Limitations and Possibilities for Future Development.** *Knowledge Organization*, 29(3/4). 198-203. 11 refs.

ABSTRACT: A translation memory system is a new type of human language technology (HLT) tool that is gaining popularity among translators. Such tools allow translators to store previously translated texts in a type of aligned bilingual database, and to recycle relevant parts of these texts when producing new translations. Currently, these tools retrieve information from the database using superficial character string matching, which often results in poor precision and recall. This paper explains how translation memory systems work, and it considers some possible ways for introducing more sophisticated information retrieval techniques into such systems by taking syntactic and semantic similarity into account. Some of the suggested techniques are inspired by those used in other areas of HLT, and some by techniques used in information science.

D. L'Homme, M.-C. L'Homme, Ch. Lemay. (2002). **Benchmarking the Performance of Two Part-of-Speech (POS) Taggers for Terminological Purposes.** *Knowledge Organization*, 29(3/4). 204-216. 19 refs.

ABSTRACT: Part-of-speech (POS) taggers are used in an increasing number of terminology applications. However, terminologists do not know exactly how they perform on specialized texts since most POS taggers have been trained on "general" corpora, that is, corpora containing all sorts of undifferentiated texts. In this article, we evaluate the performance of two POS taggers on French and English medi-

cal texts. The taggers are TnT (a statistical tagger developed at Saarland University (Brants 2000)) and WinBrill (the Windows version of the tagger initially developed by Eric Brill (1992)). Ten extracts from medical texts were submitted to the taggers and the outputs scanned manually. Results pertain to the accuracy of tagging in terms of correctly and incorrectly tagged words. We also study the handling of unknown words from different viewpoints.

R. Fugmann (2002). **The Complementarity of Natural and Index Language in the Field of Information Supply. An overview of their specific capabilities and limitations.** *Knowledge Organization*, 29(3/4). 217-230. 28 refs.

ABSTRACT: Natural text phrasing is an indeterminate process and, thus, inherently lacks representational predictability. This holds true in particular in the case of general concepts and of their syntactical connectivity. Hence, natural language query phrasing and searching is an unending adventure of trial and error and, in most cases, has an unsatisfactory outcome with respect to the recall and precision ratios of the responses. Human indexing is based on knowledgeable document interpretation and aims – among other things – at introducing predictability into the representation of documents. Due to the indeterminacy of natural language text phrasing and image construction, any adequate indexing is also indeterminate in nature and therefore inherently defies any satisfactory algorithmization. But human indexing suffers from a different set of deficiencies which are absent in the processing of non-interpreted natural language. An optimally effective information system combines both types of language in such a manner that their specific strengths are preserved and their weaknesses are avoided. If the goal is a large and enduring information system for more than merely known-item searches, the expenditure for an advanced index language and its knowledgeable and careful employment is unavoidable.

This contents page may be reproduced without charge.