Guest Editorial

Evaluating Human Language Technology: General Applications to Information Access and Management

A Special Issue of *Knowledge Organization* on Evaluation of HLT

by Widad Mustafa El Hadi

University of Lille 3

Widad Mustafa El Hadi is an Assistant Professor in Information Science at the University of Lille 3, France. She teaches in the areas of Language Technology for information systems. She holds a PhD in Linguistics from the University of Lyon II, France. Her main areas of interest include terminology, thesaurus design and Natural Language Processing-based tools evaluation. She coordinated (1996-2000), in collaboration with her colleagues at the University of Lille 3, the ARC A3 evaluation project on automatic term extraction, which was supported by the AUF (Associations des Universities Francophone). She is currently in charge of two evaluation projects within the EVALDA evaluation platform (EVALDA - a joint venture between the Ministry of Research and Technology in France) and ELRA (European Language Resources and Evaluation Association, Paris, France). The focus of one of the projects is Automatic Term extraction while the focus of the other project is Machine Translation.

This special issue of *Knowledge Organization* focuses on the central role of human language technologies (HLT) in the information society, surveys the current situation and presents contributions dealing with many areas of HLT for information access purposes. The evaluation of the technologies (systems and components) and the evaluation of applications (user-oriented/usage evaluation) are addressed across various areas of HLT relevant to spoken and written language processing, also known under the name of *Natural Language Processing* (NLP). This issue examines HLT's contribution to information access, extraction and dissemination and provides a brief account of the state-of-the-art of HLT applied to information access and management. HLT for information access is not limited to textual data. Although speech recognition and understanding is in constant development (public service operators, police, telephone operators, multimodal information systems, etc.), the study and contributions are limited to written language processing for information access purposes. I will first of all give a brief account of the areas in which many authors believe the role of technology is crucial. I will secondly define the evaluation paradigm and the specific case of HLT evaluations related to information access. Finally, I will present the selected contributions to the special issue.

The Role of HLT in Information Access

The role of language technology in information access, extraction and dissemination is essential. The radical changes in the techniques of information and communication at the end of the twentieth century have had a significant effect on the function of the linguistic paradigm and its applications in all forms of communication. The introduction of new technical means has deeply changed the possibilities for the distribution of information. Many fields show the relevance of this paradigm through the various technologies that require NLP techniques, such as document and message understanding, information detection, extraction and retrieval, question and answering, cross-language information retrieval (CLIR), text summarization, filtering, and spoken document retrieval.

NLP in information retrieval (IR), seen as a classic task, involves the retrieval of relevant documents from a large repository in response to a user query and ranks these documents according to their relevance. This task is accomplished by statistical methods that select the best representative units of document content (simple words and noun-phrases or other linguistic units). These units are used to create an index that can allow direct access to the documents containing these units. Although many sophisticated retrieval methods exist, the fundamental problem is the adequate representation of content for both the documents and the queries. Adequate representation can be achieved by transforming both the documents and query representations into weighted terms derived either from the documents or indirectly through thesauri or domain maps (Strzalkowski, Lin, & Wang, 1999). It is obvious that simple term-based representations are inadequate and that phrases denoting important concepts in domain specific databases, hence, phrase extraction, is gaining momentum. As Strzalkowski, Lin, and Wang (1999), pointed out, many systems participating in the Text REtrieval Conferences (TREC) used one or another form of phrase extraction. They gave an account of the major techniques used to obtain phrases from texts. These techniques, based on NLP techniques range from generating simple collocations, statistically generated N-grams, part-of-speech tagged sequences, syntactic structures, and semantic concepts to the most advanced ones that dig out the underlying uniformity across various surface forms of expression.

The "bag-of-words" representations common to many IR systems can hardly do justice to the complexities of free unprocessed text with which the end-user has to deal with as Strzalkowski, Lin, and Wang (1999) pointed out. Even though NLP in IR is much debated (Sparck-Jones, 1999; Fugmann, 2003, in this issue) some examples show the relevance of this technology in keyword extraction for IR (Jacquemin, 1999; 2000; 2001; and Smeaton, 1999, among others). Many authors, including Sparck-Jones, agree nevertheless on the obvious role of NLP technology in information extraction, information management and knowledge management contexts (Sparck-Jones, 1999; Maybury, 2001; Bontcheva et al., 2001; Strzalkowski, 1999; Strzalkowski, Lin, & Wang, 1999; Ruge, 1999, among many others). Others think that the success of NLP technology depends on a more radical change of focus (Strzalkowski, 1999). In other words the technology would be adapted more specifically to the areas listed above. For these reasons, the last few years have seen a growing interest in HLT in general and its applications to information access and management in particular. This is why some researchers think that NLP can offer the key to building the ultimate IR systems.

Information Extraction

Information extraction is the task of filling templates or (tables) from natural language input (Mani, 2001). The proliferation of online text motivates most current works in text interpretation. Current methods generally start by identifying key artifacts in the text, such as proper names, dates, times, and locations, and then use a combination of linguistic constraints and domain knowledge, to identify the most important content of each relevant text.

Information Management

There are a number of areas in which language technologies can improve and enable information management (IM). These have been identified as: input analysis, content-based IR, information extraction, question answering, machine translation, dialogue management, user modeling, and summarization (Maybury, 2001). More have also been identified: automatic tracking and detection of emerging topics from unstructured data (text mining), information filtering, knowledge mapping and access (lexical information, language modeling), and text categorization.

Knowledge Management (KM)

Domain-specific information in World Wide Web (WWW) ontologies are used in knowledge portals in order to narrow the gap between finding knowledge in texts and providing it to the portals. HLT is used, moreover, to reduce the cost of ontology engineering. Some applications show how language technology will help in creating new knowledge from large collections of textual information (Uszkoreit, 2001). Some current work on HLT for KM applications is reported in Maybury (2001).

Knowledge Sharing

Sharing of knowledge / knowledge transfer is "[t]he process of disseminating individual experiences, information or knowledge throughout the organization to those who (might) need it" (see European KM Framework, European KM Forum, 1999). For this

activity, HLT can provide the means for associating knowledge with the relevant decisions. The technology associated with this application is called automatic relational "hyper-linking." Relational hyperlinks are different from the simple HTML hyperlinks in that they are composed of a number of named links that can be selected from a menu. HLT is used to identify and disambiguate the concepts in the documents that need to be linked. To this end, specific techniques are used, such as named entity recognition. This method has been applied in the Hypercode system of DFKI Lab (see Uszkoreit, 2001). Many contributions can be listed in the growing field of named entity recognition and its contribution to KM.

Text Summarization

Automatic text summarization is an emerging activity in information access. The goal of summarization is to take information from a source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's application need. Sparck-Jones (1997) traced back "automatic abstracting" which was first attempted in the 1950s, in the form of Luhn's auto-extracts (see Paice, 1990). The increasing volume of machine-readable text and advances in natural language processing have stimulated a new interest in automatic summarizing. In order to have efficient systems, Sparck-Jones (1997) points out the crucial need for "The *full text revolution"* which has effects on indexing. This implies a pressing need for automatic summarizing. NLP technology provides the basic resource for this revolution. In cross-language text summarization, several languages are processed, with summaries in different languages from input (Mani, 1999, p. 22).

Information Access & CLIR Technology

The problem of multilingual access to text databases can be seen as an extension of the general IR problem. The technology used to accomplish this specific type of IR is cross-language IR (CLIR). This field is at the crossroads of both machine translation (MT) and IR. The resources to be developed in order to improve CLIR range from part-of-speech taggers, syntactic analyzers, monolingual dictionaries, bilingual dictionaries (for language pairs), to lexical databases, terminological databases, monolingual corpora and aligned corpora (for language pairs). Though CLIR functions largely as a traditional IR system, it has some specific problems not shared by monolingual text retrieval (for a complete review of a CLIR process, see Grefenstette, 2000). CLIR has been a research subfield for more than a decade now. The field has engendered three major evaluation efforts: the Cross-Language Evaluation Forum (CLEF) involving many European languages, the NTCIR Asian Language Evaluation (Chinese, Japanese and Korean), and the TREC Cross Language Track which in 2001 and 2002 focused on the Arabic Language.

These many endeavours show the relevance of HLT for information access, hence, the evaluation of technologies and applications. Researchers and practitioners in the field of information are interested, by an assessment of the contribution of these technologies in order to measure the progress achieved, to compare different approaches to a given problem and to assess system usability and user satisfaction. I will develop the evaluation paradigm in the following section.

The Evaluation Paradigm

In this section we will introduce some basic evaluation concepts and the rationale of this emergent paradigm. Evaluation plays an essential role in speech and natural language processing, both for system designers and for technology users. Evaluation activities are a corollary of the quick development of NLP tools in general and of those tailored for information retrieval, information extraction and information management in particular. It thus becomes necessary to evaluate these tools on objectively based criteria in order to have a clear picture of the state-of-the-art, assess the needs in this sector and hence promote research in this specific field. Moreover, the principal aim of existing testing methods, as reported in the literature, is to come across software errors and then try to adapt them for a particular user environment. The first endeavors in this field can be traced back to MT evaluations. A lot of contributions in this field show how the definitions of the term *evaluation* itself relate to this technology. We introduce in this section the basic concepts of evaluation. Sparck-Jones and Gallier (1996) identified three types of evaluation processes: the *adequacy evaluation*, the *diagnostic evaluation* and the *progress evaluation.* The first and second types are used for comparative benchmarking. *Adequacy evaluation* aims at finding out whether a system or product is adequate to someone's needs (see Sparck-

Jones & Gallier, 1996 and King, 1996 among many others for a more detailed discussion of these issues).

This type of evaluation is conducted when thinking of acquiring a system. It may be comparative or not, and may require considerable work to identify a user's needs. A *Diagnostic Evaluation* is described by King (1996), among many others as a type of evaluation whose purpose is to discover why a system did not give the results it was expected to give. Typically performed by a researcher developing a prototype system, such an evaluation is almost exclusively concerned with functionality characteristics and will also often make use of internal metrics based on the intermediate results the system produces. Unlike the other types of evaluation distinguished in this classification, it is a glass-box evaluation.

A *Performance Evaluation* is conducted in order to measure the performance of a system in one or more specific areas. The criteria applied in performance evaluation normally fall as Sparck-Jones and Gallier (1996) explain under two major heads '*intrinsic*' and *'extrinsic,'* also known as *intrinsic evaluation* versus *extrinsic evaluation*. Intrinsic evaluation applies to the assessment of the individual components of a system, while extrinsic evaluation assesses the overall performance of the system. The first can be compared to a glass-box evaluation model, the second one to a black-box evaluation model (see below for definitions). This distinction can be compared as Hirschman and Thompson (1997) suggested, to the performance evaluation/adequacy evaluation one, where intrinsic is to extrinsic as performance evaluation is to adequacy evaluation (see also Sparck-Jones and Gallier, 1996).

A major distinction is established between *black-box* and *glass-box.* The distinction is the following: the former considers only system input-output relations without regard to the specific mechanisms by which the outputs were obtained while the latter examines the mechanisms linking input and output (Sparck-Jones, 1996, p. 26). *Glass-box* and *black-box* evaluation is a distinction "made between component-wise versus whole-system evaluation and sometimes to a less clear-cut difference between a qualitative/descriptive approach" (Hirschman & Thompson, 1997).

Obviously a black-box approach has its pros and cons. Even if it may be criticized on account of its subjective side, end-users like it because of its usefulness when comparing two or more systems which differ in all their parameter settings (Chaudiron, 2000; Cavazza, 1993). A black-box evaluation is more oriented towards system's end-users when compared to a glass-box evaluation. For the latter the test will involve analyzing the system's functioning by looking at its different components. Each component is evaluated separately in itself. Such an approach allows for spotting and understanding the causes of dysfunctional results. It is a long term process which requires access to the internal parts of the system and an understanding of the architecture and global strategy of the software. This is obviously a *developer-oriented* approach and not an *end-user* approach (Chaudiron, 2000; Cavazza, 1993).

Another common distinction in the literature is *quantitative* versus *qualitative* evaluation. The qualitative evaluation measures, as described by Sparck-Jones and Gallier (1996, pp. 61-122), are based on observation or interviewing and are broadly designed to obtain a more holistic, less reductive or fragmented view of the situation. This type of evaluation naturally fits an end-free style. Both quantitative and qualitative approaches are goal-oriented, which means they focus on discrepancies between performance results and initial system requirements. Sparck-Jones and Gallier (1996) point out how the two types of measures are deeply interwoven although different in their nature: recall is a quantitative measure of system performance while declared satisfaction is a qualitative measure (i.e., such a measure is really qualitative even if the result of applying it to a set of users results in a percentage figure).

The qualitative approach in the evaluation process is the easiest one for end users. It means giving a value judgment on how the system globally works. The dominant approach today is towards quantitative evaluations, which are considered as more objective and reproducible than the qualitative approach. The main attempt of these approaches is to translate the concepts of relevance and quality into numerical data. Statistical approaches such as MUC 2 (Message Understanding Conference) and TREC 3 are frequently used for this type of evaluation (Chaudiron, 2000). An evaluation task can combine many approaches: it is possible to define a black-box, qualitative and comparative evaluation, or a glass-box, diagnostic and progress evaluation. The essential point is the coherence of the combined approaches.

The Evaluation paradigm is basically dependent upon two major steps: (i) Creation of textual data: raw or tagged corpora and test material. A corpus-based research is part of the infrastructure for the de-

velopment of advanced language processing applications; (ii) Test and comparison of systems on similar data (Cavazza, 1993; Adda et al., 2000).

Hirschman and Thompson (1997) sketch the successes and limitations of evaluation and focus on the major role it plays for system developers, for system integrators and for consumers. One of the major contributions of the evaluation paradigm to the design of NLP systems is the development of test corpora for spoken and written language, information retrieval and machine translation. These resources can be distributed and shared by appropriate and specialized agencies like the Linguistic Data Consortium (LDC) and European Language Resources Association (ELRA), for example. In addition, there are many different evaluation conferences or workshops and institutions which have emerged as a result of this activity and are becoming popular in HLT circles (MUCs, TREC, CLEF, Amaryllis in France, the Machine Translation Evaluation Workshops, and the Spoken Language Technology Workshops (for more evaluation projects see below).

As for the limitations of current evaluation methods, as Hirschman and Thompson (1997) have reported, there has been little focus on how the user interacts with a system. Specifically, there is no performance evaluation methodology for interactive systems, and the methodologies for adequacy evaluation are difficult to apply and not widely accepted.

There is no evaluation methodology for assessing how portable systems are to new application domains. Evaluation is labour-intensive and competes in time and resources with other activities, specifically with the development of new technical approaches. In spite of the drawbacks, evaluation methodologies will continue to progress and to develop.

Evaluation Projects

Several initiatives have been conducted in Europe and in the United States. The first initiatives on HLT evaluation can be traced back to the Defence Advanced Research Projects Agency (DARPA) and the Japanese programs on Machine Translation Evaluation. The DARPA/NIST (National Institute of Standards and Technology) Projects continued on a more regular basis, particularly with the Text Retrieval Conferences (TREC) and the Message Understanding Conferences (MUC). More and more projects are gaining momentum, be they independent or part of the famous TREC Tracks (Language Engineering Track, Question & Answering (Q & A) track), or carried out within the context of the Translingual Information Detection, Extraction, and Summarization (TIDES), or Document Understanding Conference (DUC) and the Cross-Language Information Retrieval (CLIR).

Some evaluation projects have been supported within the programs of the European Commission (TEMAA, EAGLES, ISLE, SQALE, etc.) while others have been conducted within national programs (the French evaluation projects: GRACE Project for evaluating part-of-speech taggers, the AUF four Projects on evaluating Text Retrieval Systems, Alignment Technology, Terminology Acquisition System Evaluations, and Message Understanding).

The first NLP evaluation projects in Europe have been organized by the TEMAA[1] group and the EAGLES[2] group. The two main goals of the TEMAA project are to provide a framework for the evaluation of NLP tools and a partial implementation of an overall evaluation tool. This aim is shared with the EAGLES Evaluation working group. Both of them are based on the ISO 9126 standard, which is concerned primarily with the definition of quality characteristics to use in the evaluation of software products. The work undertaken by TEMAA and EAGLES is considered as an extension to the ISO 9126 standard. This standard sets six quality characteristics: functionality, reliability, usability, efficiency, maintainability and portability. The most important item in our case is *functionality*, which is defined by ISO as « A set of attributes that bear on the existence of a set of functions and their specified properties. The functions are those that satisfy stated or implied needs ». As a first step in global NLP systems the TEMAA project has focused on spelling and grammar checkers. In both types of checkers the measures of recall and precision are the same (Maegaard, 1996).

Evaluation has become so central to progress in the speech and natural language area that many specialists in NLP technologies think it should become a research area of its own. In France, the new Evaluation platform, EVALDA, is a joint venture between the Ministry of Research and Technology and ELRA (European Language Resources and Evaluation Association, Paris, France). Within the framework of this initiative, eight evaluation projects are being conducted: *ARCADE II*: campagne d'évaluation de l'alignement de corpus multilingues; *CESART*: campagne d'Evaluation de Systèmes d'Acquisition de Res-

sources Terminologiques; *CESTA :* campagne d'Evaluation de Systèmes de Traduction automatique; *Easy: E*valuation des *A*nalyseurs *Sy*ntaxiques du français; Campagne *EQueR*, Evaluation en question-réponse; Campagne *ESTER*, Evaluation de transcriptions d'émissions radio; Campagne *EvaSY*, Evaluation en synthèse vocale; and Campagne *MEDI*A, Evaluation du dialogue hors et en contexte.

The evaluation of information access tools is, of course, of a different nature. Evaluating standard IR systems concerns in general the measurement of systems performance (relevance, recall and precision). Harter (1996) has dealt with the traditional models for experimental evaluation of IR systems and mentions those used by Cleverdon (1967); Cleverdon, Mills, and Keen (1966) (described in Harter, 1996). Many articles mention these early methods of evaluation and trace the history of evaluation from Cleverdon to the current TREC. These 1966 and 1967 evaluations tested the relative effectiveness of 33 indexing languages for retrieving information. The effectiveness, as Harter pointed out, was measured in two dimensions: by the extent to which known 'relevant' documents were retrieved and by how well retrieval of 'non-relevant' documents was suppressed. Other significant evaluations were conducted by Salton (1971). TREC later used a modification of this model, in which relevance judgments were made by assessors at the National Institute of Standards and Technology (NIST) (see Tague-Sutcliffe, 1996).

To evaluate how effective the system is, some writers believe that the original user must be involved in the relevance judgment. Others believe that at least some aspects of a system can be evaluated without relevance judgments from the users. Relevance judgments in this view represent judgments of whether or not the document is about the query and so can be made by any knowledgeable person. The other view of relevance judgments is that they represent the value of the document for a particular user at a particular point in time and so can be made by the user only at that time (Tague-Sutcliffe, 1996). Recent approaches are adopted in the TREC Conferences (Harman, 1992; 1993; 1994; 1995a; 1995b).

For testing the validity standard IR systems a number of criteria have already been drawn by the IR specialists such as Cleverdon (1962), Swanson (1977; 1988), Sparck-Jones (1981), Harter (1996), Salton (1971; 1989) and the TREC evaluating groups (Harman, 1992; 1993; 1994, 1995), among others. The most relevant type of evaluations for the scope of this special issue is the TREC NLP and the different evaluation tasks it has involved since its creation. Two contributions in this special issue, report on the authors' participation in TREC Language Engineering Tracks. A historical review on evaluation of IR systems has been also accounted for in Chen's paper in this issue.

Contributions to This Issue

Ferret, Grau, Hurault-Plantet, Illouz, Jacquemin, Monceaux, Robba, and Vilnat point out the major contributions of Question Answering technology and its introduction in TREC 8 within NLP tracks. This technology, the authors argue, reveals an increasing need for more sophisticated search engines, able to retrieve the specific piece of information that could be considered as the best possible answer to the user question. The issue intersects two domains: Information Retrieval (IR) and Natural Language Processing (NLP). According to the authors, IR is improved by integrating NLP functionalities at a large scale, that is, independently of the domain, and thus necessarily having a large linguistic coverage. This integration allows the selection of the relevant passages by means of linguistics features at the syntactic or even semantic level. In "How NLP Can Improve Question Answering," the authors show, moreover, how answering open-domain factual questions, requires Natural Language processing, for refining document selection and answer identification. The system designed at LIMSI, France, QALC, participated in the Question Answering track of the TREC8, TREC9 and TREC10 evaluations. QALC performs an analysis of documents relying on multi-word term search 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 according to the syntactic form of the question. These patterns allow QALC to handle linguistic variations at the answer level. The article focuses on the gain brought by taking into account linguistic variation in documents post-selection and in matching possible answers with a question.

In "Evaluating Chinese Text Retrieval with Multilingual Queries" Chen reports on the design of a Chinese test collection with multilingual queries and the application of this test collection to evaluation of information retrieval systems. It describes the application of the test collection in CHinese Text Retrieval (CHTR) task of NTCIR Workshop 2. The run types, effective techniques, IR models, and search results are discussed. The effective indexing units, IR models, translation techniques, and query expansion for Chinese text retrieval are identified. A tool is designed to help assessors to judge relevance and to gather the events of relevance judgment. The log file created by this tool will be used to analyze the behaviours of assessors in the future, the author indicates. Moreover, the article discusses the role of search engines in information access and the incidence of the multilingual, and the multi-cultural issues for these tools; hence their evaluation from this these perspectives. The paper reports on the collaboration of East Asian countries for construction of test collections for cross-language multilingual text retrieval.

Two contributions show how NLP can offer solutions to the inadequacies of purely statistical IR methods:

Sidhom and Hassoun in "Morpho-syntactic Parsing for a Text Mining Environment: An NP Recognition Model for Knowledge Visualization and Information Retrieval" 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 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 and management of multiform e-documents (text, multimedia, audio, image) using their text descriptions.

In "From Term Variants to Research Topics," Ibekwe-San Juan and San Juan discuss the importance of NLP in scientific and technological watch (STW) tasks. They advocate the necessity of integrating NLP technologies and go beyond the mere statistical data analysis methods (co-citation analysis, co-word analysis). The authors bring in the reasons for innovative approaches and their contribution to improve the results of such tasks. They put forward a method for STW which is NLP-oriented. The method analyses texts linguistically in order to extract terms from them. It 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.

Bowker's paper, "Information Retrieval in Translation Memory Systems: Assessment of Current Limitations and Possibilities for Future Development," focuses on the contribution of translation to information processing and management. It deals with translation memory systems and highlights their role in IR and management environments. The paper is an evaluation of the current and potential usefulness of these tools for allowing translators to access relevant information. It begins by explaining how translation memories work. It then goes on to assess some of their limitations, specifically with regard to information access and retrieval, and it ends by considering possibilities for future developments that could help to optimize the usefulness of the information retrieved by these tools. In order to maximize the usefulness of translation memory systems, the author suggests the introduction of sophisticated search techniques. They range from taking into account syntactic and semantic similarities between segments, to lemmatization and thesauri incorporation into translation memories.

L'Homme, L'Homme, and Lemay, in "Benchmarking the Performance of Two Part-of-Speech (POS) Taggers for Terminological Purposes: A Users' Viewpoint," evaluate the performance of two part-of-speech taggers on specialized corpora. The taggers are TnT (a statistical tagger developed at Saarland University and WinBrill (the Windows version of the tagger initially developed by Eric Brill). Their work is motivated by the widespread use of taggers in terminology applications and the fact that terminologists do not know exactly how they perform on specialized texts since most POS taggers have been trained

on "general" corpora. The work is also motivated by the crucial role of taggers in corpus-based terminology processing and hence the choice of a good tagger is essential. The authors' claim is that even though off-the-shelf taggers have been trained with corpora of a general nature, they are reliable enough to be used for specialized texts. The authors undertook a quantitative and a qualitative evaluation of the two taggers and came to the conclusion that different taggers perform well when applied on new corpora, and even on a specialized corpus composed of extracts from medical texts. Results obtained after applying the taggers without editing their lexicons show accuracies ranging between 93.3% and 95.15%. This proves that they are reliable tools for a terminological setting.

Fugmann's article, "The Complementarity of Natural and Index Language in the Field of Information Supply: An Overview of Their Specific Capabilities and Limitations," is more descriptive than evaluative if we consider the term 'evaluation' as defined in pages 126-128, above. The article focuses on the differences between and complementarity of natural and indexing languages. It does not address Human Language Technology Evaluation *per se* beyond extractive indexing/keyword searching. As the title shows, the author's evaluation addresses the natural language itself rather than its electronic processing in text indexing and information retrieval. His contribution is a criticism of the current mainstream opinion and claims the congenital inaccuracy of the computer for understanding any human message. In his paper, the author gives a summary of his theory on document indexing. In itself, and apart from the specific scope of this issue, the problem of indexing and of the best indexing tools remains a current question, in particular for search engine designers, though it has been one of the most widely addressed issue for some decades.

To conclude, evaluation should be an essential activity to assess a system's performance. It is essential to adapt the output of current natural language technology and resources to improve IR and information extraction techniques. NLP techniques could be used directly to produce tools for these activities by creating, linguistic resources (morpho-syntactic resources, corpora), lexical databases, terminologies, thesauri, and so forth. The future of information extraction will also depend on these vital resources. NLP technology for building the information systems of the future is "unlike today's relatively crude search engines that retrieve long lists of documents of often questionable relevance" (Strzalkowski, 1999) for "[e]ven the most advanced search engines often produce results in quite unacceptable quality" as Fugmann reported in this issue. The future systems will deliver the exact information that the user is seeking and will do so with the highest precision and reliability. To accomplish this will require the systems to 'understand' both user's information needs, as well as the information they possess in their databases," (Strzalkowski, 1999, p. 15).

This overview is intended to serve as background and introduction to the contributions in this special issue of *Knowledge Organization*. Further areas involving human language technology, in both monolingual and multilingual environments, still await exploration.

Notes

1 TEMAA-A Testbed Study of Evaluation Methodologies: Authoring Aids (Maegaard 1996); Elsnet L. JE (1994). In: Proceedings of Language Engineering Convention. CNIT, La Défense, Paris, Leeann J.E).

2 EAGLES, Expert Advisory group on Language Engineering Standards. The EAGLES initiative aims at creating to establish a set of coordinated expert groups in the area of pre-normative linguistic research. With the collaboration of more than 30 research centers, industrial organizations, professional associations across the EC, the Group is concerned, among other activities, with the evaluation and assessment of linguistic data in both the natural and speech field. See European Commission, DG XIII (1994): Linguistic Research & Engineering (LRE) an Overview

References

Adda, G., Lecompte, J., Mariani, J., Paroubek, P., & Rajman, M. (2000). Les procédures de mesure automatique de l'action GRACE pour l'évaluation des assignateurs de partie du discours pour le français. In K. Chibout, J. Mariani, N. Masson, & F. Neel (Eds.), *Ressources et évaluation en ingénierie de la langue* (pp. 645-664). Duculot, Coll. Champs linguistiques, et Collection Universités Francophones (AUF).

Béguin, A., Jouis, Ch., & Mustafa El Hadi, W. (2000). Evaluation d'outils d'aide à l'extraction et à la construction automatiques de termes et de rela-

tions sémantiques. In K. Chibout, J. Mariani, N. Masson, F. Neel (Eds.), *Ressources et évaluation en ingénierie de la langue* (pp.161-179). Duculot, Coll. Champs linguistiques, et Collection Universités Francophones (AUF).

Belkin, N. J., & Croft, W. B. (1987). Retrieval techniques. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology* (pp.109-145). New York: Elsevier Science Publishers.

Blair, D.C. (1966). STAIRS Redux: Thoughts on the STAIRS Evaluation, Ten years after. *Journal of the American Society for Information Science 47*, 4-22.

Bontcheva, K., Bewster, Ch., Ciravena, F., Cunningham, H., Gutherie, L., Gaizauskas, R., & Wilks, Y. (2001). Using HLT for Acquiring, Retrieving and Publishing Knowledge in AKT : Position Paper. *Proceedings of ACL 39th Annual Meeting. Proceedings of the workshop HLT and KM, July 6 –7, 2001*, Toulouse: CNRS, IRIT.

Carol, P. (n.d.). CLEF: an Introduction. Retrieved from http://www.clef-campaign.org.

Cavazza, M. (1993). *Méthodes d'évaluation des logiciels incorporant des technologies d'informatique linguistique*. Paris: Rapport MRE-DIST.

Chaudiron, S. (2000). The Relevance of Quality Model for NLP Applications. *Proceedings of RIAO*, *Paris, 12-14 April 2000*, 1568-1577.

Chibout, K., Mariani, J., Masson, N., & Neel, F. (Eds.). (2000). *Ressources et évaluation en ingénierie de la langue*. Duculot, Coll. Champs linguistiques, et Collection Universités Francophones (AUF).

Dorr, B. J., & Oard, D. (1998). Evaluating Resources for Query Translation in Cross-Language Information Retrieval. *Proceedings of the 1st International Conference on Language Resources and Evaluation. Granada, Spain, 28-30 May 1998*, 759-764.

European KM Framework, European KM Forum IST-2000-2639305.08.01, 11extract_V03_2001-07-31_KM_Terminology_Approaches.doc.

Grefenstette, G. (Ed). (2000). *Cross-Language Information Retrieval*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Grefenstette, G. (1998). Evaluating the Adequacy of a Multilingual Transfer Dictionary for Cross-Language Information Retrieval. *Proceedings of the 1st International Conference on Language Resources and Evaluation*. *Granada, Spain, 28-30 May 1998*, 523-524.

Harman, D. (Ed.). (1995). *Overview of the Third Text REtrieval Conference (TREC-3).* Gaithersburg, MD: NIST.

Harman, D. (Ed.). (1992). Evaluation Issues in Information Retrieval. *Information*
Processing and Management, *28*(4), 439-440.

Harman, D. (Ed.). (1993). *The First Text REtrieval Conference (TREC-1). (Special Publication)*. Gaithersburg, MD: National Institute of Standards and Technology (NIST).

Harman, D. (Ed.). *The Second Text REtrieval Conference (TREC-2)*. Gaithersburg, MD: NIST.

Harmann, D. (1998). The Text Retrieval Conferences (TRECs) and the Cross-Language Track. *Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Spain, 28-30 May 1998*, 517-522.

Harmann, D., Schauble, P., & Smeaton, A. (1997). Document Retrieval. In R. Cole, J. Mariani, H. Uszkoreit, G. B. Varile, A. Zaenen, A. Zampolli, & V. Zue (Eds.), *Survey of the State of the Art in Human Language Technology* (pp. 226-229). Cambridge: Cambridge University Press and Giardini.

Harter, S. P. (1996). Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness. *Journal of the American Society for Information Science 47*, 37-49.

Hirschman, L., & Thompson, H. S. (1997). Evaluation, Overview of Evaluation in Speech and Natural Language Processing. In R. Cole, J. Mariani, H. Uszkoreit, G. B. Varile, A. Zaenen, A. Zampolli, & V. Zue (Eds.), *Survey of the State of the Art in Human Language Technology* (pp. 232-233). Cambridge: Cambridge University Press and Giardini.

Jacobs, P. (1997). Text Interpretation: Extracting Information. In R. Cole, J. Mariani, H. Uszkoreit, G. B. Varile, A. Zaenen, A. Zampolli, & V. Zue (Eds.), *Survey of the State of the Art in Human Language Technology* (pp. 230-231). Cambridge: Cambridge University Press and Giardini.

Jacquemin, Ch. (Ed). (2000). Traitement automatique des langues pour la recherche d'information. *TAL*, *41*(2), 327-332.

Jacquemin, Ch. (2001). *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, MA: MIT Press.

Jacquemin, Ch., & Tzoukermann, E. (1999). NLP for Term Variant Extraction: Synergy between Morphology, Lexicon and Syntax. In T. Strzalkowski (Ed.), *Natural Language Information Retrieval* (pp. 25-74). Dordrecht, Netherlands: Kluwer Academic.

King M. (1996). EAGLES, Workshop, University of Geneva. Retrieved from (n.d.) http://www.issco.unige.ch/projects/eagles

Krovetz, R. (2002). On the Importance of Word Sense Disambiguation for IR. *LREC 2002, Workshop Proceedings*; *Using Semantics for Information Retrieval: State of the Art and Future Research, 2 June 2002, Las Palmas de Gran Canaria*.

Littman, M. L., Dumais, S., Landauer, T., & Thomas, K. (2000). Automatic CLIR Using Latent Semantic Indexing. In G. Grefenstette (Ed.), *Cross-Language Information Retrieval* (pp. 51-62). Dordrecht, Netherlands: Kluwer Academic.

Maegaard, B. (1996). *TEMAA: A Testbed Study of Evaluation Methodologies: Authoring Aids. Final Report presented at the University of Dublin, 1996*.

Mani, I. (2001). *Automatic Summarization*. Amsterdam: John Benjamins Publishing.

Maybury, M. (2001). Human Language Technology for Knowledge Management. *Proceedings of ACL 39th Annual Meeting. Proceedings of the workshop HLT and KM, July 6 –7, 2001*. Toulouse: CNRS, IRIT.

MUC-3. (1991). *Proceedings of the Third Message Understanding Conference*. Morgan Kaufmann.

MUC-4. (1992). *Proceedings of the Fourth Message Understanding Conference*. Morgan Kaufmann.

Mustafa El Hadi, W. (in press). Human Language Technology and its Role in Information Access, Extraction and Dissemination. *Cataloguing and Classification Quarterly*.

Mustafa El Hadi, W., Timimi, I., Béguin, A., & Debrito, M. (2001). The ARC A3 Project: Terminology Acquisition Tools: Evaluation Method and Tasks. *Evaluation Methodologies for Language ad Dialogue Systems Workshop, ACL/EACL, Toulouse, 6-7 July 2001* (pp. 41-50). Toulouse: CNRS, IRIT.

Paice, C. D. (1990). Constructing literature abstracts by computer. *Information Processing and Management*, *26*(1), 171-186, quoted in Sparck-Jones, 1997.

Picchi, E., & Peters, C. (2000). CLIR: A System for Comparable Corpus Querying. In G. Grefenstette (Ed.), *Cross-Language Information Retrieval* (pp. 81-90). Dordrecht, Netherlands: Kluwer Academic Publishers.

Ruge, G. (1999). Combining Corpus Linguistics and Human Memory for Automatic Term Retrieval. In T. Strzalkowski (Ed.), *Natural Language Information Retrieval* (pp. 75-95). Dordrecht, Netherlands: Kluwer Academic.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Publishing Company.

Salton, G. (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. (1992). The State of Retrieval System Evaluation. *Information Processing and Management*, *28*(4), 441-449.

Salton, G., & Lesk, M.E. (1993). Information Analysis and Dictionary Construction. In *The Smart Retrieval System-Experiments in Automatic Document Processing* (pp. 115-142). Englewood Cliffs, NJ: Prentice Hall.

Smeaton, A. (1999). Using NLP or NLP Resources for Information Retrieval Tasks. In T. Strzalkowski (Ed.), *Natural Language Information Retrieval* (pp.99-109). Dordrecht, Netherlands: Kluwer Academic.

Sparck-Jones K., & Gallier, J.R. (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review*. Berlin: Springer.

Sparck-Jones, K. (1997). Summarization. In R. Cole, J. Mariani, H. Uszkoreit, G. B. Varile, A. Zaenen, A. Zampolli, & V. Zue (Eds.), *Survey of the State of the Art in Human Language Technology* (pp. 232-233). Cambridge: Cambridge University Press and Giardini.

Sparck-Jones, K. (1981). Retrieval Systems Test. In K. Sparck-Jones (Ed.), *Information Retrieval Experiments* (pp. 256-284). London: Butterworths.

Sparck-Jones, K. (1999). What is the Role of NLP in Text Retrieval. In T. Strzalkowski (Ed.), *Natural Language Information Retrieval* (pp. 1-21). Dordrecht, Netherlands: Kluwer Academic.

Strzalkowski, T. (Ed). (1999). *Natural Language Information Retrieval*. Dordrecht, Netherlands: Kluwer Academic.

Strzalkowski, T., Lin, F., & Wang, J. (1999). Evaluating Natural Language Processing Techniques in Information Retrieval. In T. Strzalkowski (Ed), *Natural Language Information Retrieval* (pp. 113-142). Dordrecht, Netherlands: Kluwer Academic.

Swanson, D. R. (1988). Historical Note: Information Retrieval and the Future of an Illusion. *Journal of the American Society for Information Science*, *39*, 92-98.

Swanson, D. R. (1977). Information Retrieval Evaluation as a Trial-and-Error Process. *Library Quarterly*, *47*, 128-148.

Tague-Sutcliffe, J. M. (1996). Some perspectives on the Evaluation of Information Retrieval Systems. *Journal of the American Society for Information Science*, *47*, 1-3.

Uszkoreit, H. (2001). Human Language Technology for Knowledge Management. *Proceedings of ACL 39thAnnual Meeting. Proceedings of the workshop HLT and KM, July 6-7, 2001*. Toulouse: CNRS, IRIT.

Verlardi, P., Missikoff, M., & Basili, R. (2001). Identification of Relevant Terms to Support the Construction of Domain Ontologies. *Proceedings of ACL 39th Annual Meeting. Proceedings of the workshop HLT and KM, July 6-7,2001*. Toulouse: CNRS, IRIT.

Voorhees, E. (1993). Using WordNet to Disambiguate Word Senses for Text Retrieval. *Proceedings, 16th Annual International ACM SIGIR Conference on Research and Development in IR (SIGIR'93), Pittsburgh*, 171-180.