

John I. Tait (ed.) *Charting a new course: Natural language processing and information retrieval. Essays in honour of Karen Sparck Jones*. Dordrecht: Springer

The essays in this remarkably interesting and thought-provoking book critically examine various aspects of text retrieval, information extraction, question answering and natural language processing, and their often controversial relations to one another. It is more coherent thematically than the usual *Festschrift*, centered around the work of Karen Sparck Jones, and written by very distinguished former students, colleagues and others. As her work has spanned some 40 years, anticipating many developments and defining the assumptions underlying them, this book provides a temporal perspective seldom found in similar books. It relates modern work to the earlier period in computer science when there were hardly any computers, no machine readable resources. Conceptually, linguistics was only beginning to define natural language in the Chomskyan paradigm, though formal logic and Artificial Intelligence were fairly well advanced.

KSJ has been influential in the development of statistically based information retrieval techniques which rely very little on any kind of linguistic translation into abstract symbols. Van Riejsbergen reviews the development of probabilistic methods in document retrieval. Harman reviews the history of weighting index terms for retrieving documents relevant to a given query.

One of the major issues linking many of the 14 chapters concerns words and their properties, whether they are related to other basic words or abstract primitive symbols, and whether they have internal structure. They discuss the degree to which linguistic properties need to be recognized in IR. Porter discusses automatic stemming, removal of inflectional and derivational suffixes from morphologically complex words to identify a stem, and compares his own stemmer with an earlier and different approach, which uses a far more detailed list of suffixes, including uncommon ones; he notes their technical and linguistic differences which produce remarkably similar results. Copestake and Briscoe review noun compounds, noting their productivity (up to 8 words in length), greater frequency of shorter compounds, and the problem of structural ambiguity. They discuss various disambiguation procedures and range of paraphrases available for compositional compounds, distinguishing them from lexicalized or context dependent compounds. Pulman offers a well argued solution to the long-standing controversy over the internal structure of transitive verbs. While decomposition of these verbs into primitive semantic components derives the right inferences and explains ambiguities, it wrongly predicts synonymy or a sentence and its paraphrase into more primitive terms. Pulman proposes subevent representations which avoid abstract primes, but derive the right inferences and ambiguities.

Wilks reviews the evolving relations between Artificial Intelligence and Informational Retrieval, where the reliance typical of AI of abstract semantic objects and knowledge representation, competes with the more statistically based surface representation of words standing for themselves. He sees a convergence of previously independent areas of information technology, with statistically based techniques supplemented by linguistic representations perhaps themselves derived from corpora. Jones surveys the extension of English-based IR techniques to IR in multilingual and multimedia documents, which may involve translation from one language to

another or retrieval from documents in different languages with different word structures.

Wilks and Tait analyze critically yet sympathetically KSJ's very influential dissertation on 'clumps' of related words, which also addresses the problem of disambiguating polysemous words in information retrieval. She exploits the classification inherent in Roget's Thesaurus and in definitional dictionaries, without resolving the question of whether there are semantic primes or a fixed ontology of hierarchical relations. One of KSJ's earliest and most influential contributions has been to argue for the necessity of clear and rigorous evaluation of IR techniques. Harman gives an insightful overview of recent experimental conferences devoted to evaluation, the TREC and DUC series of which KSJ was a principal organizer. These conferences have successively refined problems to be solved by competing technologies, so as to define what works and why. Robertson surveys experiments of the 1960s comparing indexing techniques applied to a group of documents on technical subjects. The results of these small-scale experiments revealed many surprising results and limitations which led to better use of test documents and retrieval techniques. A systematic overview of methods of information retrieval was edited by KSJ in 1981, a forerunner of the TREC conferences from 1991 onward. Gaizauskas and Barker discuss how realistic the TREC conference is in posing IR problems without a socially defined use. They define actual kinds of information which journalists require in first evaluating an emerging news story, then in finding appropriate background information for it. Their discussion systematically compares the needs of journalists with the QA, summarization and IR tracks of the conferences, concentrating on the fit of technology with the search context. One of the obstacles to optimal performance is natural language. Retrieval of abstract patterns becomes perfectly accurate, as Willett shows in an extension of rigorously evaluated IR techniques to databases of chemical structures.