Recent Advances in Computational Terminology

Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme (editors)
(Université Toulouse-le-Mirail, CNRS Orsay, and Université de Montréal)

Amsterdam: John Benjamins (Natural language processing series, edited by Ruslan Mitkov, volume 2), 2001, xviii+379 pp; hardbound, ISBN 1-58811-016-8, $99.00

Reviewed by
Robert Gaizauskas
University of Sheffield

This collection of papers derives from the *Proceedings of the First Workshop on Computational Terminology* (Computerm '98), held at COLING-ACL '98 in Montreal, but is a substantial revision thereof. The current volume comprises seventeen papers plus a brief introduction by the editors. The original workshop proceedings also had seventeen papers. However, seven of these original papers have disappeared, and seven new papers have taken their place. Furthermore, of the remaining papers, most have been significantly extended. Thus, this book should not be thought of as a simple reissue, in hardcover, of the workshop proceedings.

The words *Recent Advances* in the title might be taken to suggest brave strides forward in a clear-cut research program. Nothing could be further from the truth. This is an area of largely pretheoretical research, in which researchers are struggling bravely to use computational techniques to gain some foothold in dealing with the protean complexities of real lexical usage in a variety of technical domains and in a variety of applications. Consequently the book reads a bit like the conversation of the proverbial blind men feeling an elephant, each describing the part he is feeling. This is not meant to be a criticism, for probably nothing else is possible at this time, and besides, this tends to be a feature of edited collections. It does mean, however, that a reader should not come to this volume expecting to find a coherent account of the research issues and approaches in computational terminology. There should be something in here for everyone with any interest in terminology; the danger, however, is that there may not be a meal for anyone.

Classifying the work reported in this volume is not easy. I shall cluster the papers along two dimensions, a major dimension—the task or problem addressed—and a minor dimension—the intended application. This crude structuring should help to convey some notion of the scope and content of the work. In order of ascending complexity the problems addressed by papers in the collection can be characterized as (1) term extraction—the problem of extracting a list of all and only the terms from texts in a given domain, (2) synonymy detection or semantic clustering—the problem of recognizing which terms are synonyms or belong to the same semantic class or cluster, and (3) term-oriented knowledge extraction from text—the problem of building knowledge structures in technical domains, identifying the underlying conceptual entities, attributes, and relations via terminology. The principal application areas addressed by the papers are information retrieval, terminology construction and maintenance, machine translation, automatic index extraction, and automatic abstract generation.

Consider first term extraction, the most basic of the three preceding tasks. Automatic term extraction has potential application in automatic indexing, either for

back-of-book indices or for document-collection navigation, and also for compiling controlled vocabulary terminologies such as are used in, for example, medical coding applications. Several papers address this topic. Most generically, a review paper by M. Teresa Cabré Castellví, Rosa Estopà Bagot, and Jordi Vivaldi Palatresi reviews twelve current term extraction systems, including well-known systems such as LEXTER, FASTR, TERMIGHT, and TERMS, giving a brief description of each, as well as a contrastive analysis. A paper by Lee-Feng Chien and Chun-Liang Chen addresses the problem of incremental update of domain-specific Chinese term lexicons from on-line news sources. Terms are identified and allocated in real time to topic-specific lexicons corresponding to news categories, using highly efficient data structures called PAT trees. To be acceptable for a specific lexicon a term must be *complete* (have no left or right context dependency and have an internal association norm above threshold) and must be *significant* (have a relative frequency in a document collection corresponding to the target lexicon that compares favorably to its relative frequency in a general reference collection). Béatrice Daille supplies a linguistically interesting paper on relational adjectives as signals of terms in French scientific text. Contrasting, e.g., *production importante* ('significant production') with *production laitière* ('dairy production'), she argues that in the latter type of construction, such relational adjectives frequently signal terms. She goes on to describe an automatic technique for identifying such terms that is based on looking for paraphrases of the relational adjective + noun expressed as noun + prepositional phrase, where the complement of the preposition is the nominal form of the relational adjective (so, *production du lait*). A paper by Diana Maynard and Sophia Ananiadou extends their earlier work on term recognition by operationalizing two intuitions about the role of context in termhood: first, that a candidate term that has other candidate terms in its local context is more likely to be a term, and second, that a candidate term that is similar in meaning to domain-specific terms in its local context is more likely to be a term. Toru Hisamitsu and Yoshiki Niwa focus on the specific problem of extracting terms from parenthetical expressions in Japanese news wire text. In expressions of the form $A(B)$, B might or might not be an abbreviated form of A. Segmentation problems in Japanese mean that if B is not correctly recognized as an abbreviation it will be oversegmented into single characters, causing real problems for IR systems. Hisamitsu and Niwa propose a neat solution to the problem of identifying which parenthetical expressions are genuine abbreviations that is based on a combination of statistical and rule-based techniques. Finally, a paper by Hiroshi Nakagawa carefully compares two techniques for term extraction, one based on earlier work by Frantzi and Ananiadou and the other an interesting new proposal that assesses termhood according to how productive a noun in a candidate term is in occurring in many other distinct terms.

The second of the three broad problems or tasks introduced above is the problem of synonymy detection or semantic clustering. Here the problem is not just to discover terms in text, but to relate them in basic ways. Clearly this capability is significant for information retrieval, in which documents similar in meaning to a query, but differing in expression, must be retrieved. Such a capability is also relevant, however, for automatic index creation and for automatic abstracting. Again, several papers in the collection address this topic. Akiko Aizawa and Kyo Kageura propose a technique that cleverly exploits parallel Japanese-English keyword pairs associated with academic papers to build multilingual semantically related keyword clusters for use in monolingual or cross-lingual IR applications. Peter Anick proposes to use **lexical dispersion**, a measure of the extent to which a given word is used in multiple NP constructs, to identify generic concepts in retrieval results and to structure these results accordingly. Hongyan Jing and Evelyne Tzoukermann present a stimulating new

approach to a classical problem in IR. Intuition suggests that both collapsing variant morphological forms of words and distinguishing different word senses ought to improve retrieval. But previous attempts to do so, through stemming and sense disambiguation, have not led to a reliable increase in performance. The authors present an approach based on full morphological analysis, rather than stemming, and on using context vectors to represent word sense distinctions and to determine whether identical strings in the query and the document should be matched. The approach shows improvement over a more conventional model in which string identity is the only test of synonymy. Thierry Hamon and Adeline Nazarenko start with an existing lexical resource containing synonym links and a term extractor and use these to bootstrap term synonym sets by (1) analyzing each compound term in a technical corpus into a head + expansion (modifiers) and then (2) forming candidate synonyms of the compound by combining (a) synonyms of the original head with the original expansion, (b) the original head with synonyms of the original expansion, and (c) synonyms of the original head with synonyms of the original expansion. The resulting synonym sets are to be used in a document-consulting system to help users navigate complex technical documents. Adeline Nazarenko, Pierre Zweigenbaum, Benoît Habert, and Jacques Bouaud contribute a paper that describes an approach to classifying unknown words in a medical corpus into one of the eleven top-level semantic categories in the SNOMED hierarchical terminology. They parse the corpus for NPs, extract dependency relations between the words in the parsed NPs, e.g., W_1 R W_2, and construct a graph wherein words are the nodes and edges are labeled with shared contexts between the connected words: W_1 and W_3 share a context if for some R and W_2, both W_1 R W_2 and W_3 R W_2 are attested in the corpus. In this similarity graph, words whose semantic category is known from the SNOMED resource are labeled with their category, and categories are then propagated to uncategorized nodes via a voting mechanism between the uncategorized nodes' nearest neighbors. Finally, Michael Oakes and Chris Paice describe a technique for validating terms that can occur in particular semantic roles, or slots, in an information extraction–like template structure designed to capture details of scientific papers for use in generating abstracts. Starting with an initial, corpus-derived thesaurus containing domain-specific high-frequency words and multiword units (MWUs), each manually tagged with its semantic role, the MWUs are analyzed to reveal any that contain as substrings words or shorter MWUs already in the thesaurus. For such MWUs a semantic grammar rule is generated whose pattern is the MWU with the substring replaced by its semantic role and whose action is to label matching strings with the semantic role of the MWU. Such rules, which implicitly define a class of semantically equivalent terms, generalize the thesaurus beyond observed examples and are used to validate proposed slot fillers in the template.

The third problem area, and the most challenging, is that of building knowledge-rich terminologies—terminologies that contain not only terms, but attributes and relationships of the concepts denoted by the terms, frequently for use in applications requiring controlled terminologies. James Cimino contributes a paper describing the methodology employed in maintaining a large-scale knowledge-based controlled medical terminology used to encode patient data and to provide aggregation classes for a variety of applications, such as billing and decision support. In such a critical and knowledge-rich environment, terms cannot be automatically added to the terminology as a consequence of language processing. However, Cimino describes how simple language processing, together with knowledge-based reasoning, can be used to guide a terminologist in the process of, for example, adding the name of a new drug. Anne Condamines and Josette Rebeyrolle describe a corpus-driven approach to constructing a terminological knowledge base. First they use Bourigault's LEXTER to identify

candidate terms, then initiate a search for conceptual relationships among them. Taxonomies are constructed by using a fixed set of linguistic patterns to identify candidate hypernymic and meronymic binary relations. Then pairwise comparisons are made between terms in different taxonomies, and recurrent contexts in the corpus are sought in which these term pairs co-occur. If such contexts are found, a conceptual relationship is proposed, linguistic patterns are created to match the context, and these patterns applied to the corpus to identify new terms. The process is then repeated until no more relationships or terms are found. Although highly suggestive, this paper is unclear in critical places, particularly regarding which steps are carried out manually and which automatically. Finally, Ingrid Meyer presents a framework for building knowledge-rich terminological dictionaries. Her approach is firmly semiautomatic, tools being provided to assist, rather than replace, a human terminologist. The method depends on acquiring knowledge patterns, which may be lexical, grammatical, or paralinguistic (relying on, e.g., punctuation), to find **knowledge-rich contexts** from which hypernymic or other attribute or relational knowledge may be extracted. Such patterns are acquired through an iterative manual process of refinement in conjunction with a corpus.

Not fitting neatly into the above classification are two papers on bilingual term alignment for machine translation (MT). This is an important application area for terminology systems, as the translation of terminology-laden technical documents is commercial MT's bread and butter and an area in which human translators' lack of domain-specific knowledge is likely to be a bottleneck. Eric Gaussier's paper gives a general overview of issues faced in bilingual terminology extraction from a parallel corpus, particularly choices between (1) extracting terms in each language independently, then aligning terms, or (2) parsing terms in one language, then projecting, by alignment, terms onto the second language, or (3) parallel parsing. He explores an idea, referred to as **pattern affinities**, that candidate terms expressed via one syntactic pattern in one language are more likely to be rendered in the other language by some other specific syntactic pattern but shows via an implementation of this idea using the EM algorithm that results are not significantly improved. David Hull, in a very clear and convincing piece, describes a method of bilingual lexicon construction from translated sentence pairs that relies on term extraction in the source language and a probabilistic word translation model to propose term translations in the target language. Although not perfect, this model can, the author argues, lead to significant productivity gain in constructing bilingual term lexica when used in a semiautomated mode by a human terminologist.

This is a wide-ranging collection, and the editors are to be congratulated for pulling together so much interesting material. That said, there a few reproaches to be leveled at them too. First, the level of copyediting is very poor. Spelling mistakes and minor grammatical errors abound; figures and tables have incorrect captions; formulas have undefined terms. At best this is irritating; at worst it seriously impedes understanding and conveys an impression of sloppiness that undermines the reader's trust. Perhaps this falls in the crack between what the editors and the publishers feel is their responsibility. But one expects somewhat better. Second, the book would have been much more readable and more generally useful had the papers been structured into related subareas, with introductory overviews in each area setting the stage for, and comparing and contrasting, the relevant papers. As it is, the editors have opted to present the papers in alphabetical order by first author's surname, hardly the most cognitively compelling of structural principles. The editors' introduction is a step in the right direction, but a small one, and given the wide range of topics addressed, some further analysis and guidance would have been welcome. As a consequence, although this is

a book I would regret not having in my university library, it is not one I would regret not owning myself.

Robert Gaizauskas is Professor of Computer Science at the University of Sheffield. His research interests are in applied natural language processing, specifically information extraction, most recently concentrating on biomedical texts. Gaizauskas's address is Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello St., Sheffield, S1 4DP, U.K.; e-mail: R.Gaizauskas@dcs.shef.ac.uk.