Book Reviews

Spotting and Discovering Terms through Natural Language Processing

Christian Jacquemin
(University of Paris 11)

Cambridge, MA: The MIT Press, 2001,
viii+378 pp; hardbound, ISBN
0-262-10085-1, $52.95, £36.50

Reviewed by
Sophia Ananiadou
University of Salford

Christian Jacquemin's book *Spotting and Discovering Terms through Natural Language Processing* is a much needed and most welcome addition to the field of computational terminology. The central issue of this book is the in-depth examination of term variation, that is, the morphological, syntactic, or semantic transformation of multiword terms, capturing the fact that the same concept can be expressed through term variants. From an application point of view, dealing with term variation would result in more efficient indexing engines and better term extraction tools, among other applications. The book also describes in depth the author's FASTR system, a natural language processor for term variation. There are plenty of examples throughout the book and a set of metarules in the appendix for the interested reader who would like to use FASTR.

Chapter 2 provides an excellent overview and a comparison of numerous algorithms for automatic term extraction systems as well as techniques for recognizing multiword terms. Six systems are examined in detail. The author also discusses the related area of automatic indexing, the main purpose of which is to assign content descriptors to documents. He describes 11 studies of phrase indexing (indexing through multiword units). A useful comparison of the different indexers is provided (p. 113). The aim of this extended discussion of concerns, methods, and systems is to introduce the main ideas behind FASTR, which is used for both term recognition and terminological enrichment. The identification of variants of terms is crucial in Jacquemin's term-spotting technique, and because such variation occurs in corpora frequently, "it makes sense to build an indexer on a variational mechanism" (p. 115). In addition, variants represent on average 28% of multiword term occurrences (p. 219).

Since term spotting uses various linguistic features, FASTR's grammar formalism is unification-based, inspired by PATR-II, in order to represent different types of features. Information is embedded in nontyped feature structures, which include two additional facilities: disjunction and negation. Term rules in FASTR are composed of a context-free skeleton that describes the constituent structure of terms and logical constraints (feature structures) that denote the information linked to the nodes of the context-free skeleton. The morphological model of FASTR is concatenative morphology enriched with feature structures and a list of suffixes. This applies to both inflectional and derivational morphology, although in FASTR there are two ways of describing derivational morphology: dynamically (similar to inflection) and statically,

where derivational links between words and their stems are explicitly stated in the single-word lexicon. To represent the syntax of multiword terms and to cope with term variability, FASTR's formalism uses the notions of *extended domain of locality* and *lexicalization* from lexicalized tree adjoining grammars (LTAGs) (Abeillé and Rambow 2000).

Jacquemin's approach relies on metarules that exploit syntagmatic and paradigmatic information. Syntagmatically, they describe structural mappings between, say, a multiword term and its syntactic or phrasal variants. Paradigmatically, they describe lexical relationships between the words of a term and the words of its variants that participate in the mappings. These lexical relationships may be morphological or semantic in nature or both. Morphological relationships cover words that belong to the same derivational family and express the fact that such words share a common root. Semantic relationships cover words that are linked (by synonymy or antonymy, for example); that is, they capture the fact that the words have something in common semantically. Taken together, the syntagmatic and paradigmatic elements of metarules provide complementary constraints that prevent the generation of undesirable mappings.

Chapter 4 is dedicated to the metagrammar of FASTR, which operates on top of the lexical and terminological data. Metarules dynamically transform rules written for controlled terms into new rules that, in turn, can be used to extract variants from texts. The concept of metarule is inspired by the work of Harris (Harris et al. 1989) and is further influenced by generalized phase structure grammars (Gazdar et al. 1985) and feature-based LTAGs (Srinivas et al. 1994) and by work on lexically based formalisms that seek to reduce the complexity and size of the grammar. The author contrasts his use of metarules with other formalisms (p. 157). The main features are that metarules in FASTR are more generic than those in other formalisms and, since they are dedicated to term and variant extraction, they are also much simpler. Of further interest in FASTR's metarule formalism is that it contains a compiler and an interpreter of regular expressions, and this allows the output of more complex structures (for example, coordinated structures).

Term variation is explicitly described for English through four types of elementary term variation: coordination, permutation, modification/substitution, and elision. Chapter 5 provides a practical description of how to build, tune, and evaluate metarules for syntactic term variants in English that can be produced from a controlled vocabulary (this vocabulary would typically be the result of the term extraction phase of FASTR). The chapter begins by describing variations of binary terms, for example, how to link a binary compound term such as *tooth root* with a noun phrase having the same meaning, *the root of a lower premolar tooth*, by means of permutation metarules. Section 5.2 describes how to extend variations from binary to *n*-ary terms, taking into account all structural ambiguities. A detailed description is given for an example of variation involving coordination, as this is one of the most difficult phenomena to tackle. Section 5.4 illustrates how the metagrammar can be further tuned experimentally, using occurrences extracted from the corpus through paradigmatic rules. These paradigmatic rules are refined and transformed into filtering metarules (i.e., metarules augmented with constraints), which are then used by FASTR for extracting term variants. Evaluation of the different variants extracted by FASTR shows high precision and recall. The author reports poor performance in the case of elliptic variants, however, which he notes is largely due to the noncontextuality of FASTR's parsing approach (p. 169).

The recognition of morphosyntactic variants (as opposed to the syntactic variants of Chapter 5) is dealt with in Chapter 7. FASTR's formalism is extended to include metarules for morphosyntactic variations; for example, *development of mouse embryos* is considered a morphosyntactic rather than a syntactic variant of *embryonic development*,

because the adjective *embryonic* is transformed into the noun *embryo* in the variant (p. 273). Single words contain morphological links between words and their root lemma (for English, derived from the CELEX database); these links are also expressed in the metagrammar of morphosyntactic variation. Variants of binary terms are presented one morphological transformation at a time: adjective to noun variation, noun to verb variation, and so on. The analysis of these variants is important, as terminology work has concentrated mostly on nouns, leaving unexplored non-nominal terminological occurrences, which are equally important, as they capture relations between terms.

Semantic variation, described briefly in Chapter 8, concludes the study of types of term variation. FASTR's formalism is further extended to account for this type of variation. For example, *benign mouse skin tumors* can be recognized as a semantic variant of *benign neoplasms* provided that there is a semantic link between the words *tumors* and *neoplasms* and that the insertion of the modifier *mouse skin* is accepted (p. 299). The semantic links used for extracting semantic variants in English are the synonymy links of WordNet 1.6. Morphosyntactic term variation can be seen as a special case of semantic variation, since legitimate morphological links can be established between semantically related words.

I found it more useful to read Chapter 6 after Chapters 5, 7, and 8. It describes how FASTR can be used for "incremental term enrichment," taking into account existing terms and trying to relate newly acquired terms to the previously existing terminologies. The idea is that, by deconstructing term variants, we can detect possible term associations and acquire new terms. "The variant *uterine and carotid artery* of *uterine artery* is the opportunity for discovering the term *carotid artery*. The context of acquisition shows that both terms can be coordinated, indicating that the meanings of *artery* in the original term *uterine artery* and in the candidate term *carotid artery* are similar: a blood vessel" (p. 241). The analysis of such variants produces a set of patterns, called *pattern extractors*. Each pattern extractor is then attached to a specific metarule. Starting from any variant detected by such a metarule, the extractor outputs the corresponding candidate term. This technique of term enrichment is also very useful for producing conceptual relations that can be used for automatic thesaurus construction.

In conclusion, this book is clearly written and explores an interesting and very applicable area of computational terminology: term variation. An in-depth analysis of current techniques in terminology, a detailed bibliography, and a plethora of examples demonstrating how to write rules for and use FASTR make this book an important acquisition not only for researchers interested in computational terminology but also for the wider natural language processing community. Anyone concerned with issues of automatic indexing, automatic thesaurus construction, rapid adaptation to new domains, and robust processing of not only domain-specific terminology in the classical sense but also the phrases in which variants of terms are found will learn much from this book.

References

Abeillé, Anne and Owen Rambow, editors. 2000. *Tree Adjoining Grammars: Formalisms, Linguistic Analysis and Processing*. CSLI, Stanford, CA.

Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum, and Ivan A. Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press.

Harris, Zellig S., Michael Gottfried, Thomas Ryckman, Paul Mattick, Jr., Anne Daladier, T. N. Harris, and S. Harris. 1989. *The Form of Information in Science: Analysis of Immunology Sublanguage* (Boston Studies in the Philosophy of Science, volume 104). Kluwer Academic.

Srinivas, B., Dania Egedi, Christine Doran, and Tilman Becker. 1994. Lexicalization and grammar development. In Harald Trost, editor, *Proceedings of KONVENS '94*, Vienna, pages 310–319.

Sophia Ananiadou is a Senior Lecturer in Computer Science at the University of Salford, U.K. She teaches natural language processing, and her main research interests are in computational terminology, ontologies, and information extraction. Currently, she is applying natural language processing techniques to bioinformatics. Ananiadou's address is: Computer Science, School of Sciences, University of Salford, Salford M5 4WT, U.K.; e-mail: S.Ananiadou@salford.ac.uk.