ELSEVIER

Artificial Intelligence 130 (2001) 185–189

Artificial Intelligence

www.elsevier.com/locate/artint

Book Review

Review of Natural Language Processing in R.A. Wilson and F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* ☆

Bonnie Jean Dorr

University of Maryland, Department of Computer Science, A.V. Williams Building, College Park, MD 20742, USA

The MIT Encyclopedia of the Cognitive Sciences (MITECS) is the first encyclopedia in cognitive sciences—a web-navigable resource with invaluable information and several hundred links to related resources. The material provided therein is thorough and very clearly presented by the leading scientists in each area. This is one of the most comprehensive resources in cognitive science to date. It will serve as a teaching and research guide that users may frequently refer to for important definitions, background information, and citations to relevant literature.

This review covers areas relevant to Natural Language Processing (NLP), in particular, the entries entitled "Natural Language Processing" (James Allen), "Computational Linguistics" (Aravind Joshi), "Generation" and "Machine Translation" (both by Eduard Hovy), "Computational Lexicons" (James Pustejovsky), and "Statistical Techniques" (Eugene Charniak). I will also address issues concerning the use of MITECS as an online, web-navigable document.

1. A review of natural language processing entries in MITECS

James Allen's entry entitled "Natural Language Processing" focuses on text processing —leaving speech-based processing to other entries elsewhere in the encyclopedia. The goal in NLP most relevant to cognitive science, Allen says, is that of understanding how language comprehension and generation occurs in humans. Indeed, the testing of linguistic theories is one motivation for building explicit computational models of language processing. Another, Allen says, is that of developing automatic systems for applications, e.g., automated telephone operators or web searching.

☆ MIT Press, Cambridge, MA, 1999. CD-ROM. Price US$ 149.95. ISBN 0-262-73124-X. 1312 pages. Price US$ 149.95 (Cloth). ISBN 0-262-23200-6.

E-mail address: bonnie@cs.umd.edu (B.J. Dorr).

0004-3702/01/$ – see front matter
PII: S0004-3702(01)00096-0

This entry covers three general approaches to language processing: statistical (discussed in more depth in another entry covered below), pattern-based (following traditional linguistic models, covered by other entries such as "Formal Grammar"), and reasoning-based (covered in entries such as "Planning" and "Knowledge Representation"). The applications described in this entry are categorized under the headings of Information Extraction and Retrieval, Machine Translation, and Human-Machine Interfaces. These three categories are described adequately, although it would be helpful if they were accompanied by a (mouse-clickable) table that placed each application into one of the three categories. Such a table would indicate that, for example, "Spoken-Language Dialogue Agents" falls under the heading of Human-Machine Interfaces and "Text Document Search" falls under the heading of Information Extraction and Retrieval. It should also include additional applications (not mentioned in the entry) that cut across Allen's categories, e.g., "Cross-Language Information Retrieval". (A missing citation here is [11].)

It is interesting to compare this entry to the one called "Computational Linguistics" (CL) by Aravind Joshi—a title taken by the author to be synonymous with "Natural Language Processing". Having two entries instead of one is an ingenious way of obtaining a description by two leading scientists in the field. (Both entries have links to each other.) CL is described as an interdisciplinary endeavor, at the cross-section of artificial intelligence, logic, and psychology. This entry appears to be more applications-oriented than its NLP counterpart, with a large list of applications given, including interfaces to databases, text processing, message understanding, multilingual interfaces as aids for foreign language correspondences, web pages, and speech-to-speech translation.

The entry for CL is somewhat less broad than the one for NLP, but deeper in its treatment of two areas that are mentioned only in passing in Allen's entry:

(1) Grammars and Parsers;

(2) Statistical Approaches.

I found the detailed description of grammars to be a bit of a surprise for such a general entry and wondered if certain parts of this text, such as the part on context-free grammars, might not be better described in its own entry. Joshi relies on the user's access to the 1996 Survey of State of Art in Human Language Technology (HLT) [1] for filling in the remaining details, but he cites this as a "forthcoming" reference, even though it has been out for four years. The following link should be added: http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html. The entries on Natural Language Generation (NLG) and Machine Translation (MT), both written by Eduard Hovy, are among the most thorough that I have seen. Virtually all of the key players in both fields are included. (I tried to trip up this section by searching on some of the more recent researchers such as Elhadad, but all of them were included—with the possible exception of those who presented in the most recent AMTA conferences, e.g., Nizar Habash and David Traum, see [3] below.)

In the NLG entry, Hovy describes the full architecture envisioned by leading researchers in the field (the two traditional modules of text planner and sentence realizer, with the new intermediate sentence planner to "fill the gap"), despite that this architecture has not yet been modeled in its entirety. The delightful aspect of this description is its all-inclusiveness—every researcher in NLG can be found under the heading of one or more modules included in this architecture. Details of each module are described, and some of

the more difficult issues and techniques are presented, e.g., stylistic control and the use of canned items, templates, cascaded patterns, and features.

The MT entry is equally comprehensive. All three approaches (direct, transfer, and interlingual) are adequately defined, and their three applications (assimilation, dissemination, and interaction) are clear. In addition, this is one of the very few computational entries with a discussion about evaluation, complete with pointers to the work of several researchers who are dedicated to this effort. Finally, the links to MT systems at the end of the article are very useful.

Despite its completeness, the MT entry has a few inaccuracies and omissions. The first line of the entry refers to the "fiftieth anniversary" in 1997, yet the second paragraph says MT dates back to work done in 1955, which may be confusing to readers. (Perhaps we should just call it the 55th anniversary now, in the year 2000.) The author estimates the number of worldwide companies to be 50, which I would expect to be on the low side, especially given the recent resurgence of NLP companies, i.e., "The Dot Com's" that Hovy himself has spoken about at AMTA-2000. (Note: The last six years of AMTA proceedings are missing from the bibliography of the MT entry: AMTA-1996, AMTA-1998, AMTA-2000.)

A more technical omission concerns the discussion of the use of interlinguas on a large scale. Hovy states, correctly I think, that such systems require a great deal of effort to build. However, there should be some mention of resource acquisition in this section. That is, it should be possible to build an interlingual system if lexicons existed on a large scale. Several researchers are, indeed, doing this cross-linguistically, and those researchers should be mentioned, e.g., [4]. The number of lexical items in such systems is not several hundred, as Hovy says, but tens of thousands.

This leads us to the topic of another entry, that of "Computational Lexicons", written by James Pustejovsky. The basics of a lexical entry are covered with great clarity: syntactic information (category and subcategory) and semantic information (base semantic typing and selection typing). Next, the global structure of the lexicon is described, and example of which is Pustejovsky's qualia structure (although few details are provided). Finally, a short description of the lexicon as a knowledge base is given.

Although this entry is clearly written, the presentation leaves out many details and there are several missing references, e.g., the work of Allen Cruse [2], Charles Fillmore [6], Ray Jackendoff [7], Roger Schank [13], and various researchers who have computational implementations based on these frameworks. In addition, the sub-field of lexical acquisition is missing from the description, although the bibliography identifies several readings corresponding to advances in this area (the work of Briscoe, Copestake, Wilks, etc.). This might even be a third bullet (new trends) listed at the beginning of the entry. Finally, a discussion of how one might scale up theoretical representations such as qualia (a product of the author himself) for largescale computational applications would be welcome.

"Statistical Techniques" have been used for solving a wide range of problems in natural language processing. In this MITECS entry, Eugene Charniak chooses to focus on those that fall under the heading of cognitive science, specifically: speech, part-of-speech tagging, syntactic parsing, word-sense disambiguation, and machine translation.

Areas that are excluded are lexicography and document retrieval—as the author says—as well as optical character recognition and spelling correction.

Included in this entry is the comparison between statistical and non-statistical techniques, emphasizing the depth/breadth tradeoff. In addition, the author provides a fairly clear description of the notion of Hidden Markov Model (HMM), but I found myself wondering if the description would be better filled out with a definition of the "noisy channel model". One mathematical equation would do the trick, i.e., that the most probable word given some observation can be computed by multiplying the "prior probability of w" by the "likelihood of the observation given w". Statistical machine translation, a very clear application of the "noisy channel model", is included at the end of the entry, with pointers to the relevant sections elsewhere in the encyclopedia.

2. MITECS as a web-navigable document

I was pleased by the comprehensiveness of MITECS as a web-navigable document; it was difficult to find many omissions. The MITECS search engine is a delight to use, providing a link to topics covered by individual entries. Just specifying a topic or an author associated with a topic brings the user right to the appropriate entry (or list of entries). For example, a search on the word "corpus" directs the user immediately to a list of entries that are right on the mark, including Natural Language Processing and Statistical Techniques in Natural Language Processing. Similarly, a search on a prominent name in the field, such as "Jackendoff", results in a list of expected topic areas—in this case, Semantics, Thematic Roles, Grammatical Relations, and Compositionality.

A possible enhancement to the document would be the provision of a mechanism for updating the pages—so that later editions could be more easily produced. For example, there are already a number of broken or changed links (e.g., "Le Journal" on the main NLP page). In addition, several of the entries have bibliographic references that end as early as 1994; there have been many advances in the late 90's that are not included in these entries. This has resulted in some glaring omissions, including the lack of reference to Jurafsky and Martin's Speech and Language Processing textbook [5], which is arguably the most influential NLP book of the last half decade. Additional researchers with recent work also are not linked to any entry in the encyclopedia, e.g., Judith Klavans [8], Dekang Lin [9], Kathy McKeown [10], and Martha Palmer [12]. In at least one of these cases (McKeown), it is likely there is a software bug in the search engine since I found a citation in Hovy's entry on "Natural Language Generation". Updating the web links and bibliographic references upon the arrival of new, seminal works, would be a big plus for this online document.

Another possible improvement would be a comprehensive index of names of people in the field with links to relevant sections in the text—including bibliographic references (stored separately). Although there is an author index for individual entries in the encyclopedia, there is no list of names mentioned in the text of these entries. In an attempt to find descriptions of researchers' works in different areas of natural language processing, I used the MITECS-provided search engine—but this was a trial-and-error effort that often resulted in searches for names that do not appear anywhere in the document. It would be easier if the full, alphabetized list of names were available for searching and navigating.

Despite these (very minor) suggestions for improvement, I view MITECS as the most comprehensive and useful document currently existing for instruction and research. In addition to clear textual descriptions and critical citations, each of the 471 entries contains links to numerous additional online resources including course syllabi, FAQs on different topics, links to organizations and newsletters such as ACL, AAAI, and Colibri, and homepages for NLP software such as Ask Jeeves. I intend to use MITECS extensively over the next several years and will make parts of the volume required reading for students in my Natural Language Process and Artificial Intelligence classes.

References

[1] R.A. Cole (Ed.), Survey of the State of the Art in Human Language Technology, NSF, European Commission, 1996.
[2] D.A. Cruse, Lexical Semantics, Cambridge University Press, Cambridge, 1986.
[3] B.J. Dorr, N. Habash, D. Traum, A thematic hierarchy for efficient generation from lexical-conceptual structure, in: Proc. 3rd Conference of the Association for MT in the America's, Langhorne, PA, 1998, pp. 333–343.
[4] B.J. Dorr, Large-scale dictionary construction for foreign language tutoring and interlingual machine translation, Machine Translation 12 (4) (1997) 271–322.
[5] D. Jurafsky, J.H. Martin, Speech and Language Processing, Prentice-Hall, Upper Saddle River, NJ, 2000.
[6] C. Fillmore, The case for case, in: E. Bach, R. Harms (Eds.), Universals in Linguistic Theory, Holt, Rinehart, and Winston, 1968, pp. 1–88.
[7] R. Jackendoff, Semantics and Cognition, MIT Press, Cambridge, MA, 1983.
[8] J. Klavans, P. Resnik, The Balancing Act: Combining Symbolic and Statistical Approaches to Language, MIT Press, Cambridge, MA, 1997.
[9] D. Lin, Automatic identification of non-compositional phrases, in: Proc. ACL-99, University of Maryland, College Park, MD, 1999, pp. 317–324.
[10] K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay, E. Eskin, Towards multidocument summarization by reformulation: Progress and prospects, in: Proc. AAAI-99, Orlando, FL, 1999.
[11] D. Oard, A. Diekema, Cross-language information retrieval, Annual Review of Information Science and Technology 33 (1998) 223–256.
[12] M. Palmer, H.T. Dang, J. Rosenzweig, Sense tagging the penn treebank, in: Proc. 2nd Language Resources and Evaluation Conference, Athens, Greece, June 1–4, 2000.
[13] R.C. Schank, Conceptual Information Processing, Elsevier Science, Amsterdam, 1975.