



Recent Advances in Natural Language Processing

Fabio Ciravegna, *University of Sheffield*

Sanda Harabagiu, *University of Texas, Dallas*

Language is the most natural way of communication for humans. The vast majority of information is stored or passed in *natural language* (for example, in textual format or in dialogues). *Natural language processing* aims at defining methodologies, models, and systems able to cope with NL, to either understand or generate it.

Opportunities for NLP applications range from querying archives (for example, the historical activity on querying databases), to accessing collections of texts and extracting information, to report generation, to machine translation.

Ongoing progress

The past few years have witnessed several major NLP advances, allowing text and speech processing to make vast inroads in universal access to information.

It is a common experience nowadays to contact call centers that use speech-understanding systems (for example, when accessing travel information). We will see even more of this in the future. One main field of application is the ever-growing World Wide Web. Owing to recent progress in open-domain question answering, multidocument summarization, and information extraction, in the near future we'll be able to explore information posted on the Web, such as is often in large online document collections, fast and efficiently. Instead of users having to scan long lists of documents deemed relevant to a Boolean query (as when using search engines), they will receive a concise, exact answer to their questions expressed in natural language. Similarly, multidocument-summarization techniques will help solve the current information overload by producing short abstracts that coherently combine essential information scattered among multiple documents. Moreover, research efforts are enabling question answering, summarization, and information extraction in a wide range of languages—

for example, English, Spanish, and Italian, and even Arabic, Farsi, Urdu, and Somali.

NLP is a multidisciplinary field at the intersection of linguistics (classic and computational), psycholinguistics, computer science and engineering, machine learning, and statistics. Recent advances have only been possible thanks to developments in each of these fields, from linguistic-resources definition (for example, large corpora), to computational models, to evaluation exercises. Large strategic funds have been devoted to NLP in the last few years, from the DARPA TIPSTER program in the 90s and the current Advanced Research and Development Activity's ACQUAINT (www.ic-arda.org/InfoExploit/aquaint) and DARPA TIDES (www.darpa.mil/iao/TIDES.htm) programs in the US, to the European funds for linguistic engineering in the 90s and the current Information Society Technology (www.cordis.lu/ist) program.

As usual, advances in research work in concatenation; small advances allow big technological jumps, which in turn allow other advances. The availability of large resources such as the Penn Treebank¹ (www.cis.upenn.edu/~treebank) has made possible the development of large-coverage, accurate syntactic parsers.^{2,3} Semantic knowledge bases such as WordNet⁴ and, more recently, FrameNet⁵ have moved semantic processing from the back burner, enabling exciting applications such as question answering on open-domain collections. Semantic parsers⁶ are under development, allowing deeper processing of language. In parallel, the emerging field of text mining allows computational linguists

to customize text or speech processing on the basis of the most salient relations.

Additionally, numerous speech and text databases available from the Linguistic Data Consortium (www ldc.upenn.edu) have let more researchers than ever share data and compare techniques. Corpus-based NLP has let us discover in the past decade how to exploit the statistical properties of text and speech databases. With considerable interest focused on the new de facto corpus for modern NLP procedures and the availability of the Semantic Web standards, the field is set for an exciting new research and development phase.

In this issue

This special issue's theme focuses on recent NLP advances. The seven articles specifically identify the state of the art of NLP in terms of research and development activities, current issues, and future directions.

A special issue of a magazine can, of course, allow only a partial representation of the current development in a field. However, we believe that the articles in this issue are largely representative of at least some NLP major trends. There are articles describing NLP-based applications and articles analyzing language methodologies, providing, we believe, the right mix of technological insight into the state of the art.

"Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays," by Jill Burstein, Daniel Marcu, and Kevin Knight, describes a decision-based discourse analyzer using decision-tree machine learning algorithms combined with boosting techniques for obtaining an optimal discourse model. This model is based on a rich feature set, comprising rhetorical-structure-theory relations, discourse markers, and lexico-syntactic information. This discourse analysis software, embedded in the Criterion writing-analysis tools, helps students develop sophisticated essays by accessing knowledge of discourse structure.

"Speaking the Users' Languages," by Amy Isard, Jon Oberlander, Ion Androutsopoulos, and Colin Matheson, presents a system that generates descriptions of unseen objects in text of various degrees of complexity. It tailors these descriptions to the user's sophistication—for example, for adults, children, or experts. Moreover, the system can generate descriptions in English, Greek, or Italian.

"Ontology Learning and Its Application to Automated Terminology Translation," by Roberto Navigli, Paola Velardi, and Aldo Gangemi, depicts a system that automatically learns ontologies from texts in a given domain. The

domain-specific ontologies that result from these knowledge acquisition methods can be incorporated into WordNet, a lexico-semantic database that numerous NLP systems employ.

"Personalizing Web Publishing via Information Extraction," by Roberto Basili, Alessandro Moschitti, Maria Teresa Pazienza, and Fabio Massimo Zanzotto, portrays a knowledge-based information extraction approach that enables multilingual text classification as well as hyperlinking. This approach generates hyperlinks and their motivations on the basis of similarity between events extracted from document pairs.

"Machine and Human Performance for Single and Multidocument Summarization," by Judith Schlesinger, John Conroy, Mary Ellen Okurowski, and Dianne O'Leary, surveys modern summarization techniques and their evaluations at the 2002 Document Understanding Conference.

"Intelligent Indexing of Crime Scene Photographs," by Katerina Pastra, Horacio Saggion, and Yorick Wilks, introduces a novel indexing approach that exploits predicate-argument relations and textual fragments. Additionally, their approach extracts possible relations for better indexing and uses a crime domain ontology.

"Automatic Ontology-Based Knowledge Extraction from Web Documents," by Harith Alani, Sanghee Kim, David Millard, Mark Weal, Wendy Hall, Paul Lewis, and Nigel Shadbolt, presents Artequakt, a system that automatically extracts knowledge about artists from the Web to populate a knowledge base and to generate narrative biographies. ■

Acknowledgments

Organizing a special issue is both difficult and tiring. We thank all those who submitted papers for this special issue. Space limitations and the necessity of representing a wide range of technologies caused the exclusion of some other excellent papers. They will be published in future issues of this magazine. Finally, we thank Pauline Hosillos for taking care of the everyday work for this special issue.

References

1. M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational Linguistics*, vol. 20, no. 2, 1993, pp. 313–330. Reprinted in *Using Large Corpora*, Susan Armstrong, ed., MIT Press, Cambridge, Mass., 1994, pp. 273–290.
2. M. Collins, "A New Statistical Parser Based on Bigram Lexical Dependencies," *Proc. 34th Ann. Meeting Assoc. for Computational Linguistics (ACL 96)*, 1996, pp. 184–191.
3. E. Charniak, "A Maximum Entropy-Inspired

The Authors



Fabio Ciravegna is a senior research scientist at the University of Sheffield's Department of Computer Science. His research interests include adaptive information extraction from text. He is Sheffield's

technical manager for Advanced Knowledge Technologies, a project dealing with knowledge-based technologies for advanced knowledge management. He is also the coordinator of the dot.com consortium, a project dealing with adaptive information extraction for knowledge management and the Semantic Web. Finally, he is a co-investigator in the MIAKT (Medical Informatics AKT) project. Before joining Sheffield, he was principal investigator and coordinator for information extraction at the Fiat Research Center and at ITC-Irst. Contact him at the Natural Language Processing Group, Dept. of Computer Science, Univ. of Sheffield, Regent Court, 211 Portobello St., Sheffield, S1 4DP, UK; f.ciravegna@dcs.shef.ac.uk; www.dcs.shef.ac.uk/~fabio.



Sanda Harabagiu is an associate professor and the Erik Jonsson School Research Initiation Chair at the University of Texas, Dallas. She also directs the university's Human Language Technology

Research Institute. Her research interests include natural language processing, knowledge processing, and AI, and particularly textual question answering, reference resolution, and textual cohesion and coherence. She received her PhD in computer engineering from the University of Southern California and her PhD in computer science from the University of Rome Tor Vergata. She has received the NSF Career Award. She is a member of the IEEE Computer Society, AAAI, and Association for Computational Linguistics. Contact her at the Dept. of Computer Science, Univ. of Texas at Dallas, Richardson, TX 75083-0688; sanda@cs.utdallas.edu; www.utdallas.edu/~sanda.

Parser," *Proc. 6th Applied Natural Lang. Conf. and 1st Meeting North Am. Chapter of the Assoc. for Computational Linguistics (ANLP-NAACL 2000)*, Morgan Kaufmann, San Francisco, 2000, Section 2, pp. 132–139.

4. C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, Mass., 1998.
5. C.R. Johnson and C.J. Fillmore, "The FrameNet Tagset for Frame-Semantic and Syntactic Coding of Predicate Argument Structure," *Proc. 6th Applied Natural Lang. Conf. and 1st Meeting North Am. Chapter of the Assoc. for Computational Linguistics (ANLP-NAACL 2000)*, Morgan Kaufmann, San Francisco, 2000, Section 2, pp. 56–62.
6. D. Gildea, and D. Jurasky, "Automatic Labeling of Semantic Roles," *Computational Linguistics*, vol. 28, no. 3, Sept. 2002, pp. 245–288.