Current Developments in Natural Language Processing at Apple

Branimir Boguraev, Apple Computer, Inc. & Christopher Kennedy, University of California, Santa Cruz

Over the past several years, Apple Computer, Inc. has been actively pursuing the goal of developing its natural language program — both in terms of establishing a core set of NLP technologies, and in defining application areas for these. This development is under the direction of Dr. Branimir Boguraev, formerly of IBM's T.J. Watson Research Centre and the Computer Laboratory at Cambridge University. Organizationally, the natural language work is carried out at the Apple Research Laboratories; it is positioned within the Intelligent Systems Program (Dr. James Miller) in the Knowledge Management Group, and is coordinated with some on-going speech work (Dr. Kim Silverman) and information access research (Dr. Dan Rose).

Projects under Boguraev's supervision investigate a range of issues, including:

- optimal packaging of a substrate of NLP functionalities, with appropriate Application Programming Interfaces (API's), embedded within the Macintosh Operating System (Mac OS);
- their synergistic integration with other information technologies;
- studies of how NLP can be leveraged for further enhancing the user experience; *and*
- building several information management systems incorporating linguistic processing of text-based documents.

Under Apple's internship program, Boguraev works with graduate students in Linguistics and Computer Science, assisting in the development of prototype systems; Apple also participates, jointly with the Computer Science Department at Brandeis University, in NSF-sponsored projects within the Human Language Technology program, looking at some more practical aspects of current research in computational lexical semantics. Especially within Apple, the emphasis has been on finding suitable tasks within which to embed linguistic functionalities; on striking the right balance of scalable and robust technologies which can reliably analyze significantly large text sources; and on developing algorithms for focused semantic analysis starting from a relatively shallow syntactic base.

A case in point is technical terminology identification. Traditionally, this has had somewhat limited use, primarily for indexing purposes; at Apple, however, it has been applied, within a general domain acquisition framework, to the task of instantiating databases for instructional assistance. In particular, an NLP environment has been customised to derive help databases automatically, by parsing on-line software manuals. This work was largely carried out as a joint effort with Michael Johnston (previously with the Linguistics Department at the University of California at Santa Cruz, and now at the Oregon Graduate Institute) and Scott Waterman (formerly of Brandeis University, now with Price Waterhouse Research Laboratory). A more recent project brought together Chris Kennedy (Linguistics, University of California at Santa Cruz) and Marc Verhagen (Computer Science, Brandeis University); the focus here is on semantically-driven content analysis of arbitrary texts, and exemplifies certain aspects of the more algorithmically-oriented work at Apple.

On-line help and domain functionality

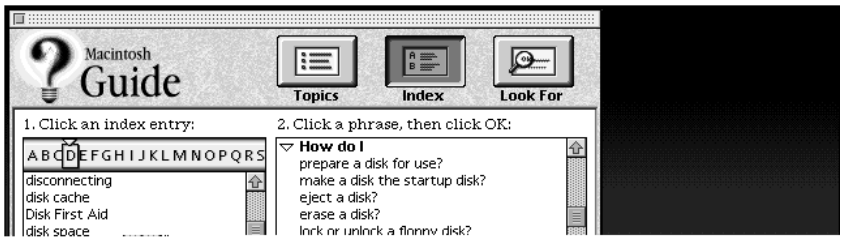
Apple Guide is an integral component of the Macintosh operating system; it is a general framework for on-line delivery of context-sensitive, task-specific assistance across the entire range of software applications running under the Mac OS. The underlying metaphor is that of answering user questions like *"What is X?"*, *"How do I do Y?", "Why doesn't Z work?" or "If I want to know more about W, what else should I learn?"*. Answers are 'pre-compiled', on the basis of a full domain description defining the functionality of a given application. For each application, assuming the existence of such a description in a certain database format, Apple Guide coaches the user through a sequence of definition panels (describing key domain concepts), action steps (unfolding the correct sequence of actions required to perform a task or achieve a goal), or cross-reference information (revealing additional relevant data concerning the original user query).

An Apple Guide database is typically instantiated by hand, on a per application basis, by trained instructional designers. Viewed abstractly, however, the information in such a database constitutes complete domain characterization for the application—in terms of domain objects, their properties, and relations among them. In order to answer questions like those above, certain aspects of the domain (in this case that of operating system level activities) need to be identified: a kind of a domain object is a disk; there are several types of disk (including floppy disks, start-up disks, and internal hard disks); disks need to be prepared for use; floppy disks can be ejected; and so forth.

It is clear that a terminology identification component, applied to suitably chosen technical documentation for the domain, could be profitably utilized for the purposes of such domain characterization. We have found that the core set of terms from a technical document can be refined not only to include all (and only) the domain objects, but also to enrich the descriptions of these domain objects, by deriving relational structures for each of them (not unlike generating lexical entries for a dynamically induced lexicon). This type of staged lexical acquisition ultimately derives a conceptual map of the technical domain. Mapping from such a domain description to an Apple Guide database is relatively straightforward. Terms are 'place-holders' for definitions of salient domain objects. Relations naturally map onto a *"How do I ...?"* panel. In database format terms, this would mean definition

entries for *"start-up disk", "network connection", "System folder"*; and action sequence panels for *"How do I specify a start-up disk?", "How do I use the internal hard disk as a start-up disk?"*, and so forth. The definitions and task sequences would still have to be supplied externally, but the generation of the database is fully automatic.

The outcome of such a domain acquisition and mapping process is illustrated below.



One of the screen snapshots is from the 'canonical' Macintosh Guide for the standard Mac OS configuration; the Guide database here is built, manually, by a team of instructional designers. The other snapshot displays, through the same delivery mechanism, a database which has been fully automatically constructed, by the system outlined above, following an analysis of the primary technical documentation for this domain (Macintosh User's Guide). Note, in particular, the *"How do I..."* leading to detailed instructions concerning common tasks with specific objects (in this example, disks) in the MacOS domain. Barring nonessential differences, there is a strong overlap between the two lists (*"prepare a disk for use", "eject a disk", "test (and repair) a disk", "protect a file/information on disk"*, and so forth). Moreover, some additional action types have been identified, which are clearly relevant to this domain, but missing from the 'canonical' database: *"share a disk", "find items on a disk"*.

Content analysis and document characterisation

More recent work addresses the problem of identifying the core content-bearing units in arbitrary texts, with a focus on smaller documents (in comparison with the technical prose discussed in the previous section), and allowing for wide diversity of genre. This work is centred around the development of a set of sophisticated text processing tools based on a very shallow linguistic analysis of the input stream, in which depth of base level analysis is traded off for breadth of coverage (the analysis engine currently used is Lingsoft's supertagger). Two complementary lines of attack here are exemplified by Verhagen's work on identification, extraction, classification and typing of proper names, technical terms, and other complex nominals from text, which extends the core phrasal analysis engine originally developed for term analysis, and Kennedy's work on algorithms for topic-based segmentation of text, salience-driven anaphora resolution, and content characterisation.

The work on anaphora resolution provides a representative illustration of our goal of deriving high-level semantic analysis from an impoverished input stream. The basic strategy we employ is a modified implementation of an algorithm developed by Lappin and Leass, which relies heavily on fully parsed inputs. We have found that by combining the phrasal analysis generated by an extended implementation of the term identification technology with an analysis of the overall topical structure of the text (derived by comparing adjacent blocks of text for overall lexical similarity, after Hearst), we can achieve precision of resolution comparable to that of Lappin and Leass' algorithm. Our algorithm, together with its implementation and detailed evaluation, is presented in a COLING-96 paper.

Like the Lappin and Leass algorithm, our anaphora resolution procedure determines the *salience* of all referential expressions in a text. Roughly speaking, salience is a measure of the relative prominence of an object in a discourse: objects with high salience are the focus of attention; those with low salience are at the periphery. This measure is not only an important factor in determining the antecedent of a pronoun; it also provides a basis for establishing a partial ordering on a term set, which may then be used as the basis for a characterisation of a document's content in terms of those expressions which identify the most prominent participants in the discourse. Ongoing work at Apple is aimed at using this reduced set of terms (in combination with relational information of the sort discussed in the previous section and packaged in an appropriate presentation metaphor) as the basis for a highly representative and meaningful—but at the same time compact—document abstraction.

FOR INFORMATION
Branimir Boguraev can be contacted at:
Apple Research Laboratories, Apple Computer, Inc.
One Infinite Loop, MS: 301-3S
Cupertino CA 95014, USA
Tel: +1 408 974 1048
Fax: +1 408 974 8414
Email: bkb@research.apple.com

Christopher Kennedy can be contacted at:
Department of Linguistics, Stevenson College
University of California at Santa Cruz
Santa Cruz CA 95064, USA
Tel: +1 408 459 4765
Fax: +1 408 459 3334
Email: kennedy@ling.ucsc.edu