Panel on Corpus Linguistics and Information Retrieval

Robert Krovetz
Computer Science Department
University of Massachusetts, Amherst, MA 01003

Corpus Linguistics is becoming increasingly important. The most recent conferences of the Association for Computational Linguistics (ACL), and the International Conference on Computational Linguistics (COLING) include a large number of papers about corpus analysis. Archives are being developed for storing large bodies of text, and lexical databases are beginning to be shared and analyzed. This panel will discuss some of the recent results in this area, and the bearing they have on Information Retrieval. Some of the areas of overlap include:

1. Determination of text type (generally via various types of lexical distribution). This can give a coarse analysis that would be useful for automatic assignment of index terms, word sense identification, automatic assignment of a text to a specialized database, and other applications.

2. Use of a tagged corpus. Recent research in Information Retrieval has shown that a tagged corpus can improve retrieval performance via the identification of phrases.

3. Comparisons of corpora and machine-readable dictionaries. This can help give us a better understanding of overall coverage, and how texts differ with respect to sublanguage.

4. Statistical parsing based on a corpus of parsed text.

5. Research on word collocations. This is an important issue in Information Retrieval with respect to proximity search in large text databases. It also bears on the use of 'statistical phrases', and trying to determine useful retrieval units above the word level.

6. Word sense disambiguation. Research has shown that word senses provide a significant separation between relevant and nonrelevant documents. However, it is an open question as to whether disambiguation can make an improvement in the performance of a retrieval system.

Most of the research on Corpus Linguistics has been done in Europe, primarily in England, the Netherlands, and Scandanavia. The International Computer Archive of Modern English (ICAME) is located in Bergen, Norway, and is one of the main resources for researchers in this area. Corpus Linguistics has an annual conference (always held in Europe), and there is almost no overlap in the literature between Corpus Linguistics and Information Retrieval. One of the main purposes of this panel is to make the two communities aware of each other's work. Both communities focus primarily on large text databases, and the statistical analysis of the information they contain. The Corpus Linguistics community differs from most of Computational Linguistics in its focus on large amounts of real-world data, and in its emphasis on statistical analysis.

The panel members are: Willem Meijs, Roger Garside, Yves Chiaramella, and Kenneth Church. Willem Meijs is a member of the Alpha Informatics Department of Amsterdam University, and is the editor of many of the proceedings of the Corpus Linguistics conferences; his research is on the extraction of semantic information from machine-readable dictionaries. Roger Garside is a lecturer in the Department of Computing at the University of Lancaster, and co-director (with Professor G.N. Leech) of UCREL (Unit for Computer Research on the English Language); his research is on fast methods for the annotation of corpora, and on methods for stochastic parsing of the text. Yves Chiaramella is a Professor in the Computer Science Department of the University Joseph Fournier of Grenoble and director of the Laboratorie de Genie Informatique (Laboratory for Computer Science Engineering); his research is focused on the use of natural language processing in Information Retrieval. Kenneth Church is a senior researcher at AT&T Bell Laboratories, and has been conducting

[0]

research on word collocations and word sense disambiguation. Each panelist is giving a short presentation which will be followed by a general discussion. The abstracts of these presentations are given below.

Text Anotation and Stochastic Parsing

Roger Garside
Department of Computing
Lancaster University

Interest in corpus linguistics has grown enormously in the last five years in the Natural Language Processing community. Corpora, originally seen only as providing a resource of information on language usage for the exclusive use of linguists, are now used for lexicography (eg the CoBuild dictionary, the British National Corpus Project); in the training of speech recognition and text-to-speech systems; in the evaluation of grammars; and in the development of grammars, whether by probabilistic training of a non-probabilistic grammar or by direct induction of some sort of probabilistic grammar.

Corpora can be annotated in various ways (with the provenance of the various segments, or the prosody, etc) but some of the most interesting are annotated with syntactic and/or semantic information. There are several procedures which can be used to annotate a corpus, depending on the complexity of the annotation required. A corpus can be annotated automatically by some natural language processing system, possibly followed by a manual post-editing phase if the accuracy of the system is not high enough; it can be annotated in an interactive way, with the computer and the manual analyst sharing the annotation task (typically with the analyst called on to make a decision at each point of ambiguity); or the process can be wholly manually, but with a special-purpose editor to speed the input of the annotation and to check its validity. The choice between these procedures depends on the accuracy of any available automatic method, and crucially on the comparative speed of correction versus input by hand. There is a need for consistency, accuracy and speed in the annotation, and this will limit the level of detail which can be achieved in the annotation scheme.

The easiest type of syntactic annotation is word-tagging; ie the assignment of a part-of-speech mark to each word of a corpus, the parts of speech being taken from a pre-defined tagset. There are a number of automatic tagging systems in existence for English and various other languages (often making use of Hidden Markov Model mechanisms), and these can assign tags with a percentage accuracy rate in the high 90s. For additional accuracy, manual post-editing can be applied. Word-tagging is usually a preliminary step to any of the more complex annotation systems.

For more complete syntactic annotation, there is a problem in the non-availability of automatic parsing systems capable of robust and complete parsing of naturally occurring text with a sufficiently high accuracy rate. Two approaches have been taken to this. At the University of Pennsylvania Mitch Marcus has processed text through the Fidditch parser, and has a team of analysts correct the output by hand. The Fidditch parser produces a highly-accurate but incomplete parse (for example, it leaves many prepositional phrases unattached), so the analysts would complete the parse, for example by attaching the prepositional phrase at the appropriate place.

At Lancaster we have taken a different approach. After an initial attempt to have analysts manually insert a full parse for each sentence, according to a simple but complete grammar, and having investigating the speed at which this could be done, we settled on the mechanism of skeleton parsing. Here the analysts insert labelled or unlabelled brackets into sentences which have been previously word-tagged, bracketing only those word groups which have uncontroversial internal cohesion. This procedure marks the main constituents in a conventional parse tree, with the rationale that this skeleton parse would enable a training mechanism to extract probability figures for a grammar attempting to generate a full parse (this being the primary aim of the generation of the "treebank" of parsed sentences). The advantage of this method is its speed (the peak annotation rate is of the order of a sentence per minute), but also that the resultant annotation is as theory- neutral as possible.

Some work has also been done at Lancaster on rudimentary semantic annotation (word-disambiguation, anaphor resolution) and this will be reported on if there is time.

Relating Dictionaries and Corpora

Willem Meijs
Alpha Informatics Department
Amsterdam University

The study of machine-readable dictionaries (MRDs) makes it possible to trace the inherent relational network implicit in dictionary-definitions in great detail. By combining centrally-related definitions in MRDs for the major parts of speech it is possible to trace the relations implicit in MRDs automatically and thus link them up, so that they combine into taxonomies. This is what we have done at Amsterdam University in the 'LINKS' project for the Longman Dictionary of Contemporary English (LDOCE), and we are applying similar techniques to a number of Van Dale dictionaries in the context of the ESPRIT project 'ACQUILEX'. In 'ACQUILEX' we are now turning taxonomies into knowledge representations in the form of typed feature-hierarchies which allow (default) inheritance, logical inferences and a certain amount of common-sense reasoning. In an obvious sense the information in any dictionary - including ones that are not, like the Cobuild dictionary, in fact based on a corpus - is intended to reflect the way in which people use the words that figure as entries in them. Considerations having to do with available space, manpower, money etc. restrict the number of examples (if any) to just a few supposedly characteristic illustrative sentences or phrases. This often has the adverse effect of fixing semantic representation too much at the level of the individual word, where a 'collocationally sensitive' representation might be much more appropriate. Given present-day capacities for mass storage and retrieval in electronic environments, however, we can overcome this restriction by relating MRD information systema- tically to corpus data, in effect using the MRDs as index and access media to up-to-date (and ideally updatable) corpora with statistically significant, representative data. My view is that the best way to do this is in terms of a General Lexicon (GL) indexed to a 'vanilla-type' General Corpus (GC), and a number of Specialized Lexicons (SLs) indexed to Specialized Corpora (SCs). The GL can be constructed on the basis of a general 'commoncore' dictionary in MRD-form. With the words (or senses, rather) in the GL belonging to specific domains we move into 'expert' areas. Since a general-purpose dictionary caters for the general reader or learner, the definitions for these senses in the MRD will reflect that, and they may thus not always be quite adequate from the experts' point of view. The important thing is that these areas in principle provide natural contact-points to pass to and fro between the GL and any number of a SLs that can be constructed on the basis of material such as special-purpose dictionaries, manuals, guides etc. that may be (made) available in machine-readable form. The advantage of linking SLs and the GL up via such transition-points is that in this way the expert knowledge is not isolated from the general knowledge, but is linked up with it in a natural way: the more specific hierarchies and taxonomies in the various expert domains are thus automatically embedded in the more general ones contained in the GL. The various SLs should each be indexed to an associated Specialized Corpus (SC) consisting of texts in their specific domains. In order to make this kind of setup work we need computationally simple but powerful (and fast!) software that can scan the available corpus-material and relate it in sensible ways with the MRD-derived relational networks. The use of parallel processing may make it possible to cope with the massive pattern- matching and statistical computation which such an approach would require, given the fact that 'large corpora' nowadays are in the order of a hundred million words or more rather than the one- million words that used to be the standard size for 'first- generation' corpora like BROWN and LOB.

Using Bilingual Materials to Develop Word Sense Disambiguation Methods

Kenneth W. Church
AT&T Bell Laboratories

Word sense disambiguation has been recognized as a major problem in natural language processing research for over forty years. Both quantitive and qualitative methods have been tried, but much of this work has been stymied by difficulties in acquiring appropriate lexical resources, such as semantic networks and annotated corpora. In particular, much of the work on qualitative methods has had to focus on "toy" domains since currently available semantic networks generally lack broad coverage. Similarly, much of the work on quantitative methods has had to depend on small amounts of hand-labeled text for testing and training.

We have achieved considerable progress recently by taking advantage of a new source of testing and training materials. Rather than depending on small amounts of hand-labeled text, we have been making use of relatively large amounts of parallel text, text such as the Canadian Hansards, which are available in multiple languages. The translation can often be used in lieu of hand-labeling. For example, consider the polysemous word *sentence*, which has two major senses: (1) a judicial sentence, and (2), a syntactic sentence. We can collect a number of sense (1) examples by extracting instances that are translated as *peine*, and we can collect a number of sense (2) examples by extracting instances that are translated as *phrase*. In this way, we have been able to acquire a considerable amount of testing and training material for developing and testing our disambiguation algorithms.

The availability of this testing and training material has enabled us to develop quantitative disambiguation methods that achieve 92 percent accuracy in discriminating between two very distinct senses of a noun such as *sentence*. In the training phase, we collect a number of instances of each sense of the polysemous noun. Then in the testing phase, we are given a new instance of the noun, and are asked to assign the instance to one of the senses. We attempt to answer this question by comparing the context of the unknown instance with contexts of known instances using a Bayesian argument that has been applied successfully in related tasks such as author identification and information retrieval.

The Use of Corpus Linguistics Techniques in the IOTA Project

Yves Chiaramella
Department of Computer Science
University Joseph Fourier of Grenoble

The use of NLP techniques seems mandatory to enhance indexing languages for textual documents and for improving user interfaces. The goal is usually to recognise more elaborate concepts within texts and thus to define more representative semantic models for documents and queries. This area of research is very active and most researchers are aware that specific approaches have to be defined to cope with the particular needs and constraints in Information Retrieval Systems: the necessity to restrict the linguistic tools to manageable problems given the complexity of natural language ambiguities, and the necessity to process large amounts of texts. In other words there is no general definition of the linguistic knowledge that has to be included in IR Systems because their requirements may be extremely different considering the characteristics of the corpus and of the users. Whatever the chosen approaches, IR Systems need linguistic knowledge and we have observed in the IOTA and RIME projects that this knowledge is not easy to find. Academic definitions given by dictionnaries and linguistic literature are always somewhat incomplete when confronted by actual texts. The Corpus Linguistics approach is certainly of great help in that it has developed methods and tools for making explicit this additional knowledge. Within the IOTA project we have analyzed the co-occurrences of stemmed words, and this has allowed us to identify domain-specific knowledge given by noun-phrases, and to use those noun phrases to improve the indexing of technical documentation. We will describe our experiences in the IOTA project with word-coocurrences as well as our work with the Tresor de la Langue Francaise (Treasure of the French Language), a large Corpus Linguistics project from the late seventies.

References

Research on Corpus Linguistics is usually published in the ICAME Journal, which is published by the Norwegian Computing Centre for the Humanities in Bergen, and in Computational Linguistics, the journal of the Association for Computational Linguistics. The Corpus Linguistics conferences are published in book form and are given below. An online bibliography of the literature in Corpus Linguistics is available from ICAME (FAFSRV@NOBERGEN.BITNET).

Corpus Linguistics: Recent Advances in the Use of Computer Corpora in English Language Research, Jan Aarts and Willem Meijs (eds), Amsterdam: Editions Rodopi, 1984

Corpus Linguistics II: New Studies in the Analysis and Exploitation of Computer Corpora, Jan Aarts and Willem Meijs (eds), Amsterdam: Editions Rodopi, 1986

Corpus Linguistics and Beyond, Willem Meijs (ed), Amsterdam: Editions Rodopi, 1987

Corpus Linguistics: Hard and Soft, Merja Kyto, Ossi Ihalainen, and Matti Rissanen (eds), Amsterdam: Editions Rodopi, 1988

Theory and Practice in Corpus Linguistics, Jan Aarts and Willem Meijs (eds), Amsterdam: Editions Rodopi, 1990

English Computer Corpora: Selected Papers and Research Guide, Stig Johansson and Anna-Brita Stenstrom (eds), Berlin: Mouton de Gruyter, 1991