0.0.1 Data-Intensive Models

Data-intensive models – that is, models which make extensive use of empirical knowledge obtained by analysing large corpora – have played an increasingly important role in computational linguistics over the last decade. Corpora range from highly specialised examples, as typified by the Edinburgh MapTask corpus of marked-up dialogs, to more general, balanced, corpora such as the Penn Treebank and Saarbrücken's Negra corpus. The wide variety of available corpora provides researchers with substantial "real" language data concerning language use. This is turn enables the improvement of both the practical coverage of models, while also forging a closer relationship between computational models and observed human linguistic performance.

From an engineering perspective, large corpora provide both a target for the performance of real language processing systems, as well as a source of data for training current probabilistic systems. Such probabilistic models range from those which attempt to achieve rich analysis, accuracy, and broad coverage to more practically-oriented techniques aimed at solving restricted problems. Uszkoreit's group in Saarbrücken uses shallow processing techniques, which gain their value from being trained on large quantities of text, to give useful analyses relatively inexpensively. Information extraction, for example, is one of the most active and promising fields in language technology. Since full understanding of unrestricted texts will remain unfeasible for practical deployment for quite some time, IE technology provides the means for recognising and analysing selected relevant pieces of information from NL texts.

From a cognitive perspective, large corpora have long been seen as an important way of estimating lexical biases. More recently, however, psychological models have begun to include more sophisticated probabilistic mechanisms, which rely on corpora as approximations of "human linguistic experience". This further opens the door for a synergy between probabilistic language technology and cognitive models, as explored in Crocker's recent research. While cognitive models have traditionally focussed on explaining a relatively small number of "interesting" pathological phenomena, it is important that these models be expanded to account for human processing of normal, real language. For example, Steedman's research on the induction of CCG grammars from corpora naturally complements Crocker's work on the development on broad-coverage probabilistic models of human sentence processing.

Finally, it is worth mentioning that in recent years a third research area has started to emerge, namely the development of "hybrid systems" which combine the best of data-intensive and symbolic approaches. In most of the other areas, discussed in the following sections, there is a pervasive data-intensive thread, which highlights the increasing importance of this dimension to most areas of language research.

Pinkal's groups in Saarbrücken is also exploring the use of corpora as a basis for developing computational lexica, and also to drive development of more practically oriented semantic processing and dialog models. Relevant Edinburgh work in this area includes Taylor's research on the induction and use of probabilistic dialog models for spoken language understanding. Webber's interests in information retrieval, particularly in the medical/health domain, complements the activities of Uszkoreit's group. In addition, Williams' research on machine learning and probabilistic modeling contributes a foundational support for many of the topics relevant to this theme.

Example Thesis Topic: Accurate Crosslingual Information Extraction. Supervised by Uszkoreit (Saarbrücken) and Webber (Edinburgh).

In this thesis project, existing IE systems from the LTG at the University of Edinburgh and from DFKI Saarbrücken will be used as the basis for crosslingual extensions as well as for the combination of shallow and deep analysis methods. The LTG in Edinburgh has built crosslingual IE components that can be extended to support German. The LT Lab at DFKI has systems in crosslingual information retrieval and efficient deep analysis technology.

Thus the thesis addresses two problems: (i) crosslinguality, and (ii) improvement of accuracy through selected utilisation of deep analysis methods. In crosslingual access, analysis takes part in the language of the document, whereas the results are presented in the language of the information customer. Crosslingual IE is especially useful in international corporations and organisations. The selected application of deep analysis methods is meant to help in situations where the shallow analysis can identify relevant phrases and relations but is unable to assign the phrases unambiguously to the roles of the relations.

Additional Thesis Topics

- Clustering techniques for smoothing probabilistic models of selectional restrictions (Crocker/Williams)
- Improving key-sentence summarization by recognizing and exploiting source text discourse connections (Webber/Uszkoreit)
- Investigations of whether human parsing decisions are based on frequency (Pickering/Crocker)
- Combining shallow and deep methods for prosody generation in Text to Speech Systems (Klein/Barry)
- A Data-Intensive Strategy for Computing Discourse Structure (Lascarides/Pinkal)