

0.0.1 Lexicon, Ontology, Knowledge Representation

For computational linguists, knowledge (of syntax, semantics, phonology, morphology, and much else besides) ultimately stems from the lexicon. A key topic in computational linguistics is how to structure and organize the (vast) amount of information in lexica of realistic size.

In AI, on the other hand, knowledge is stored in ontologies, information about how concepts are hierarchically related. Substantial attention has been devoted to developing techniques for building and maintaining ontologies, and for extracting and reasoning with the knowledge they contain (for example, the development and implementation of so-called description logics).

These two enterprises have much in common. For a start, lexica and ontologies for real applications are large objects, and the techniques required to build and maintain are similar. Moreover, standard repositories of lexical information (notably WordNet and EuroWordNet) are in essence large ontologies structured with additional information (such as which two concepts are opposites).

The research we envisage here lies at the border of lexicon and ontology construction. On the one hand, we are interested in constructing lexica structured to take advantage of current work in ontology (such as research in description logics). On the other hand, we are interested in developing linguistic systems for interfacing with large existing ontologies (such as large-scale mathematics ontologies).

Both Saarbrücken and Edinburgh have experience in these areas. In Saarbrücken, Pinkal is involved with the development of large semantic lexica for language technology applications; Blackburn with logics for knowledge representation; and Kohlhasse with mathematical knowledge bases and ontology. In Edinburgh, Steedman has developed a sophisticated temporal ontology for representing and reasoning about states and events, and Lascarides' work deals with modeling lexical interpretation.

This area is of considerable importance to the proposed EGK because only by working with large scale lexica and knowledge bases can one learn why certain methods will (or will not) scale up. Thus it might be helpful to think of this area as adding an empirical dimension to the proposed work on “Inference” and “Dialogue and Generation”: it is here that we experiment with the applications which tell us about the feasibility of various approaches.

Example Thesis Topic: Lexical Knowledge and Description Logic.

Supervised by Blackburn (Saarbrücken) and Lascarides (Edinburgh).

Existing lexical knowledge bases such as WordNet provide substantial amount of information — but in a rather inflexible form. To make use of this information in natural language systems we need to be able to use this information in a variety of ways: for example we need to be able to reason about the compatibility or incompatibility of various concepts, to find the most specific lexical item that captures some required idea, and to ensure that lexical choices are compatible with our knowledge of the situation.

Modern description logics provide both the theoretical background for analyzing such tasks and efficiently implemented tools for actually carrying them about. Because large lexica are similar to large ontologies, this suggests using description logic both to encode and extract information. On a practical level, viewing a lexicon not simply as a database, but as an ABox (the description logic terminology for information about particular individuals) of lexical entries, coupled with a TBox (the description logic terminology for theoretical generalizations) it should be possible to extract appropriate lexical information in a more flexible way. On a theoretical level, the use of description logic to construct lexicons – and in particular, the use of

the TBox to distribute generalizations over all lexical entries – suggests that computational lexicons and world knowledge can be integrated in a natural way. The proposed thesis would explore both the practical and theoretical consequences of the move to description logic oriented approaches. The work would draw on Blackburn’s work on hybrid and description logics, and Lascaride’s work on linguistic applications of default information.

Additional Thesis Topics

- Building a Mathematical Knowledge Base by Controlled NL (Kohlhase/Klein)
- Consistency maintenance in a large-scale semantic lexicon covering heterogeneous domains (Pinkal/Moore)
- The semantics of temporal adverbials (Blackburn/Steedman)