Modelling semantics: developing a cognitively plausible, data-driven approach.

Ann Copestake

1. Background

Research on compositional semantics has largely progressed in line with work on syntax. Reasonably broad coverage, `deep', grammars have been created for a range of languages that use logical forms as the output from analysis and sometimes also as the input to generation (realisation). Although computational compositional semantics is hardly a solved problem, there is a considerable degree of consensus about the main approaches: for instance, it is now standard to work with logical forms which are underspecified for quantifier scope. The situation with lexical semantics is very different: there is little agreement on theoretical representation or even the scope of the field (the debate on lexical versus encyclopedic knowledge keeps recurring). Out of necessity, most researchers interested in broad coverage experiments use WordNet as a resource. There has been much emphasis on word sense disambiguation, generally with respect to WordNet senses. But WordNet has essentially adopted the sense listing convention, following the practice of printed dictionaries. This is not cognitively plausible and may have limited practical utility, especially given the difficulty human annotators experience in assigning senses.

There seems to be general (though not universal) agreement that a cognitively plausible account of word meaning representations should involve some idea of the semantic `space' that a word occupies relative to other words. Such a meaning representation allows some word senses to be conceptualised as corresponding to relatively distinct pockets of meaning (homonyms: e.g., the classic `bank' senses), while other senses occupy spaces which are less distinct (e.g., `twist'): see e.g., Rodd et al (2004). There has been surprisingly little computational work on deriving sense distinctions automatically from corpora, though see Schütze (1997).

Corpus-based approaches to deriving non-symbolic semantic representations depend on producing huge vectors of word co-occurrences. Although work along these lines has proved useful for applications that require text categorisation or classification, it mostly ignores the structural information that is given by syntax and also
the consistent patterns of word sense distinctions, such as metonymy. Exceptions are work by Lin (see, e.g., Pantel and Lin, 2002) and Pado and Lapata (2003) where models are constructed over syntactically parsed
data: this is the sort of line that we propose to pursue.

2. Overview of proposal

The proposal here is to develop a hybrid representation, taking account of compositional properties and regular relationships between word uses. For instance, a ditransitive verb should be thought of as involving at least four semantic spaces: one for each subcategorised

syntactic argument and one for the event itself. Shallow syntactic processing (e.g., RASP: Briscoe and Carroll, 2002) has progressed to a point where it is possible to do large scale experiments along these lines and underspecified semantic representations can be constructed from RASP parses (Copestake, 2003). Such an approach relies on a very tight integration of statistical and symbolic components: statistical techniques are not just used for choosing between symbolic analyses but are a fully integrated component of the model.

To be cognitively plausible, such models should support the classical lexical semantic relationships of hyponymy, meronymy and perhaps antonymy, at least in the cases where human judgements give clear results. However, we cannot expect a simple spatial inclusion effect, if the lexical semantic representations is derived from co-occurrences in text: for instance, the fact that basic level categories are used in preference to other terms complicates the results.

This approach would treat word senses via some form of soft clustering (cf., Schutze, 1997 and Neill, 2003). However, regular relationships between senses (e.g., conventionalised metonymy, verb alternation) cannot be ignored. Such effects must either be modelled symbolically (e.g., using lexical rules) or as analogies: currently it is not clear which approach is more cognitively plausible.

Clearly, semantics cannot be studied in isolation from discourse context. Work on metonymy by Lapata et al (2003) and Frisson and Pickering (in press) is clearly relevant to the programme being proposed here: plausible semantic models must allow for further specification in a discourse context and possibly also for overriding of default interpretations.

3. Preliminary notes on specific tasks

Model development:

Initial work could be done using existing techniques for shallow parsing and underspecified compositional semantics on text corpora. There are a wide range of statistical and machine learning approaches that might be applied. Work should be carried out on multiple languages, but this would involve further development of robust parsing tools.

Data gathering:

Current large scale corpora include samples from many genres and contexts, but are predominantly text rather than speech. This may cause problems in developing cognitively plausible models since they do not reflect an individual's experience of words and word uses. We should investigate how seriously this biases our models by acquiring a series of relatively small scale corpora that reflect the language that particular individuals are exposed to over a time period (i.e., longitudinal corpora, cf., Campbell (2004)) and comparing them to the large scale corpora.

Evaluation:

Considerable work will be required to develop an appropriate evaluation methodology. However, a simple word prediction task might be very useful, since it is cognitively plausible and is much simpler to implement than embedding the model inside an application such as question answering. Positive results could also have a direct practical impact, for instance in systems for augmentative and alternative communication.

4. Collaboration:

Expertise to make interdisciplinary progress exists in the UK, including (minimally):

Cambridge: Marslen-Wilson et al, Copestake, Briscoe.
Edinburgh: Pickering, Keller, Lascarides.
Sheffield: Lapata.

5. References:

Briscoe, Ted and John Carroll (2002)
`Robust accurate statistical annotation of general text'
3rd International Conference on Language Resources and Evaluation (LREC-2002),
Las Palmas, Gran Canaria

Campbell, Nick (2004)
`Speech and Expression: the value of a longitudinal corpus'
4th International Conference on Language Resources and Evaluation (LREC-2004),
Lisbon

Copestake, Ann (2003)
`Report on the design of RMRS'
Project deliverable, DeepThought --- Hybrid Deep and Shallow Methods for Knowledge-Intensive Information Extraction

Frisson, S. P. and Pickering, M. (in press)
`The Processing of metonymy: Evidence from eye movements'
Journal of Experimental Psychology, Learning, Memory, and Cognition

Mirella Lapata, Frank Keller, Christoph Scheepers (2003)
`Intra-sentential context effects on the interpretation of logical metonymy'
Cognitive Science, 27:4, 651-670.

Neill, Daniel (2002)
`Exploring Semantic Classification'
MPhil thesis, University of Cambridge

Sebastian Pado and Mirella Lapata (2003)
`Constructing Semantic Space Models from Parsed Corpora'
In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,
pp 126--135, Sapporo, Japan.

Patrick Pantel and Dekang Lin. 2002. Discovering Word Senses from Text. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery

and Data Mining 2002. pp. 613-619. Edmonton, Canada.

Rodd, Jennifer M., Gaskell, Gareth M., Marslen-Wilson, William D. (2004)
`Modelling the effects of semantic ambiguity in word recognition' Cognitive Science, 28, 89-104

Hinrich Schütze (1997), `Ambiguity Resolution in Language Learning: Computational and Cognitive Models', CSLI Publications