From Semantics to (Plausible) Inference

1 Semantics for Underspecified (R)MRS

We have seen that it is possible to construct an underspecified semantic representation of sentence meaning compositionally in (R)MRS. However, although much of this representation is motivated by work on formal semantics (e.g. generalized quantifiers), (R)MRS itself is not a logic with proof and model theory. Rather it describes sets of trees of well-formed formulas in a neo-Davidsonian version of FOL extended with generalized quantifiers. This implies that if you want to do inference and actual interpretation then it is still necessary to expand out the set of formulas and work with these. For instance, given the input (1a), a parser should produce a mostly resolved (R)MRS like (1b).

(1) a Every man loves some woman

b l1:every(x, h1, h2), l2:man(x), l3:love(e), l3:arg(e, x), l3:arg2(e, y), l4: some(y, h3 h4), l5:woman(y), h2=$_q$ l3

c every(x man(x), (some y, woman(y), love(e), arg1(e, x), arg2(e, y)))

d some(y, woman(y), every(x man(x), love(e), arg1(e, x), arg2(e, y)))

From (1b) we can create two fully specified formuli (1c) or (1d). Given an appropriate model and theorem prover we can then compute truth-values or reason that (1d) entails (1c), etc. However, we can't do this directly with (1b). For some tasks this may not matter; e.g. for (S)MT we might be able to generate directly from (1b) into another language which also underspecifies quantifier scope morphosyntactically (most do).

Koller and Lascarides (2009) provide a model theory for RMRS which captures how removing underspecification reduces the set of trees of logical

formuli denoted by a RMRS. This lays the groundwork for defining satisfiability of RMRSs and an entailment relation between RMRSs. This takes us a step closer to being able to reason directly with RMRS representations.

2 Boxer

Bos (2005, 2008) has developed the approach to obtaining a wide-coverage FOL semantics from CCG to support reasoning. Firstly, he uses Discourse Representation Theory (DRT) as his semantic representation. This is very similar to MRS in that it is a neo-Davidsonian FOL with generalized quantifiers and a similar approach to conjunction of formuli which was historically developed to handle anaphora better, rather than to support (more) underspecification; e.g. in (2a) and (2b), the pronouns function semantically like bound variables within the scope of *every* and *a*:

(2) a Every farmer who owns a donkey beats it.
b Every farmer owns a donkey. He beats it.
c every(x, farmer(x), some(y, donkey(y), own(x y), beat(x y)))

That is the (simplified) semantics of these examples is captured by (2c). For (2b) it is fairly easy to see that syntax-guided traslation of sentences into FOL will lead to problems as the translation of the first sentence will 'close off' the scope of the quantifiers before the pronouns are translated. Something similar happens in (2a), at least in classical Montague-style semantics (as in Cann's book). Bos & Blackburn (2004) discuss DRT and pronouns in detail.

Although, DRT provides a technical solution that allows something similar to elementary predications being inserted into an implictly conjunctive semantic representation within the scope of quantifiers (i.e. to fill a hole / link to a hook in MRS terms), this doesn't really solve the problem of choosing the right antecedent for a pronoun. So Bos (2008) extends Boxer with a simple anaphora resolution system and Bos (2005) extends it with meaning postulates for lexical entailments derived from WordNet (see next section).

At this point, Boxer is able to output a resolved semantics for quite a large fragment of English. This can (often) be converted to FOL and fed to a theorem prover to perform inference and to a model builder to check for

consistency between meaning postulates and Boxer's output. Bos' papers give examples of inferences that are supported by the system and discuss where the system makes mistakes. The inferences mostly involve comparatively simple hyponymy, synonymy relations and the mistakes mostly involve discourse interpretation (pronouns, presuppositions). The off-the-shelf technology that he uses also means that natural, generalized quantifiers can't be handled unless they translate into FOL quantifiers. Nevertheless, the coverage of real data is unprecedented and impressive.

3 Word Meaning

Formal semantics has largely ignored word meaning except to point out that in logical formuli we need to replace a word form or lemma by an appropriate word sense (usually denoted as bold face lemma prime, lemma-number, etc (*loved*, **love′** / **love1**). We also need to know what follows from a word sense and this is usually encoded in terms of (FOL) meaning postulates:

(3) a $\forall x, y$ love′(x, y) $\rightarrow$ like′(x, y)
b $\forall x, y$ love′(x, y) $\rightarrow \neg$ hate′(x, y)
c $\neg\forall x, y$ desire′(x, y) $\rightarrow$ love′(x, y)

Although this is conceptually and representationally straightforward enough, there are at least three major issues:

1. How to get this information?
2. How to ensure it is consistent?
3. How to choose the right sense?

Bos solves 1) by pulling lexical facts from WordNet (nouns) and VerbNet – these are manually created databases (derived in part from dictionaries) which are certainly not complete and proabably inconsistent. The information they contain is specific to senses of the words defined, so is only applicable in context to a word sense, so Bos simply assumes the most frequent sense (sense 1, given Wordnet) is appropriate. If the background theory built via WordNet/VerbNet is overall inconsistent, because the data is inconsistent, the algorithm for extracting relevant meaning postulates doesn't

work perfectly, or a word sense is wrong, then the theorem prover cannot be used or will produce useless inferences.

There has been a lot of work on learning word meaning from text using distributional models of meaning (see Turney and Pantel, 2010 for a review and/or Word Meaning and Discourse Understanding Module). These models cluster words by contexts using approaches which are extensions of techniques used in information retrieval and document clustering, where a document is represented as a bag-of-words and retrieved via keywords indexed to documents, or the word-document matrix is reduced so that documents are clustered.

Words can be clustered according to their distributional similarity by choosing a representation of context (other words in a document or local window around the traget word, or set of words to which the target is linked by grammatical relations), obtaining word-context frequency counts from texts, and then clustering according to these (normalized) counts. This provides a general notion of word similarity where word senses are 'blended', to obtain a representation of word senses identified by contexts, we need to do second order clustering over the word vectors clusters at the first stage (and allow words to associate to more than one sense cluster). There are many ways to go about both steps, but one that is conceptually quite clean and results in a conditional probability distribution of word senses given a word is to use Latent Dirichlet Allocation (LDA) (as described in lecture 8 of the ML4LP Module). This is one of two approaches evaluated in Dinu and Lapata (2010) which works well. This work provides a more motivated way of picking a word sense to associate with a word occurrence in context than Bos' and so goes some way to solving 3) above.

Other researchers are trying to extend distributional semantics to recover more than just a notion of word (sense) similarity (clustering) so that the sort of information that Bos derives from WordNet/VerbNet might be learnable directly from text, but so far this work has not produced results comparable with these manual resources. So it seems that for the moment we can at best supplement these resources with some domain-specific incomplete and possibly inconsistent information using data-driven techniques.

4 Probabilistic Theorem Proving

Machine learning offers many models for classification (i.e. plausible propositional inference of the form:

$\forall x\ p(x) \wedge\ q(x) \rightarrow\ C(x)$

Probabilistic / statistical relational inference of the form, e.g:

$\forall x, y\ P(x, y) \wedge\ Q(x, y) \rightarrow\ R(x, y)$

is far less advanced. Recently, some progress has been made which is beginning to influence NLP and semantic interpretation.

Markov Logic Networks (MLNs, Richardson & Domingos, 2006) extend theorem proving to plausible probabilistic reasoning with finite (small) first-order models in a theoretically-motivated and representationally convenient way, and thus open up the possibility of reasoning in the face of partial knowledge, uncertainty and even inconsistency. Some of the inspiration for MLNs comes from NLP work on statistical parsing as the approach basically applies a maximum entropy model to FOL. Garrette *et al.*, give a succinct introduction to MLNs and then explore how they can be used in conjunction with Boxer to (partially) resolve issues 1) and 2) above. They also deploy an approach similar to Dinu and Lapata to resolve 3) above.

5 Weighted Abduction

Abduction is somewhat like (minimal) model building in that it allows the introduction of supporting premises to aid interpretation. Intuitively, abduction is reasoning from consequent to antecedent. For example, knowing (4a), (4b) and (4c), we might conclude (4d) on the basis that drunkenness is more common than fever (weights: w1 > w2).

(4) a w1, ∀ x drunk(x) → stagger(x)
b w2, ∀ x fever(x) → stagger(x)
c stagger(kim)
d drunk(kim)

This inference is not deductively valid, because we need to assume the antecedent in order to prove the consequent. In the 80s and eraly 90s Jerry Hobbs and colleagues developed a theory of language interpretation based around weighted abduction. (See section 3.1 of handout 2 from Intro to NLP for an example.) Weights were asigned manually and weight combination, especially when combining multiple sources of informaton – it is Saturday night in a club or it is Monday morning in a doctor's surgery – was dificult. Scaling the approach would require a great deal of background knowledge.

Blythe *et al.* (2011) show how the output of Boxer/uDRT can be used for weighted abduction using MLNs. This solves the weight combination problem at least and is, in principle, compatible with Bos (2008) integratin of WordNet and other resources for background knowledge with Boxer. Their system manages to make 18/22 inferences necessary to interpret a small testset of cases requiring abduction / plausible inference.

6 Quantifier Scope Resolution

Underspecifying quantifier scope is all very well but for at least some language interpretation tasks choosing the most likly scoping is necessary. Manshadi and Allen (2011) present a dataset with scopes resolved and use supervised classification to assign narrow/wide or no scope classes to pairs of quantified NPs. This allows them to scope the quantifiers in test data using features sych as quantifier type, order, head of the NP, etc. They achieve an accuracy of 78% which is better than the human annotators managed on this task.

The approach is crude compared to the techniques for computing scoped logical forms from (R)MRS and from uDRT and isn't guranteed to produce a globally consistent set of scopings. It also doesn't model the scope interactions between quantifiers and logical operators (e.g. negation), so there is a need for more sophsticated integration of scope resolution and underspecified semantic representations.

7 Question-Answering

A number of studies have used supervised machine learning techniques to learn 'semantic parsers' that map from text to logical representations, but these require training data matching sentences to logical forms which can only be produced by experts. Liang *et al.* (2011) develop a system which, given a question returns the correct answer. It learns appropriate logical representations for questions to compute answers using a domain knowledge base from a training dataset of question-answer pairs. They develop a simpler dependency-based semantics which is similar to a dependency tree representation of RMRS. This representation is convenient as it can be learnt more simply as a mapping from the output of a parser which returns a syntactic dependency represenation. They demonstrate that the resulting system is able to scope quantifiers and negation, handling ambiguities in (elliptical) comparative constructions, by learning the most likely logical forms on the basis of optimising performance on the traning data.

This is an impressive peice of work and probably the rewsulting system is the most robust and sophisticated semantic interpreter extant. However, the approach seems limited to QA for now.

8 Conclusions

It is hard to know where computational semantics / language interpretation will be in a few years' time. After languishing from the early 90s 'til recently (whilst the field pursued statistical / machine learning approaches, and ignored compositional semantics) suddenly logical semantics is back in fashion, partly because of recent advances in probabilistic logic/inference. We are still some way from robust wide-coverage language interpretation, but I expect to see fast progress over the next few years, because many of the key pieces needed to build a system are in place: wide-coverage compositional semantics, distributional semantic space models of word meaning, large knowledge bases (WordNet, FrameNet, FreeBase, Yago, etc), and better theorem provers, model builders, and probabilistic inference engines (Church, Alchemy, Tuffy).

Homework
Do the readings below for the next three lectures and come to them prepared

to ask and answer questions. We'll look at the papers on Boxer and RMRS (handout 3 and Koller/Lascarides) first, then those by Dinu and Lapata, Garrette *et al.*and Blythe *et al.*, and finally Manshadi and Allen and Liang *et al.*.

Reading

Blythe J., Hobbs, J. *et al.*, Implementing weighted abduction in Markov logic, Int. Wkshp on Computational Semantics, 2011
aclweb.org/anthology-new/W/W11/W11-0107.pdf
Dinu G, & M. Lapata, Measuring distributional similarity in context, EMNLP 2010
aclweb.org/anthology-new/D/D10/D10-1113.pdf
Garrette, D., K. Erk & R. Mooney, Integrating logical representations with probabilistic information using Markov logic, Int. Wkshp on Computational Semantics, 2011
aclweb.org/anthology-new/W/W11/W11-0112.pdf
Koller, A & A. Lascarides, A logic of semantic representations for shallow parsing, ACL 2009
aclweb.org/anthology-new/E/E09/E09-1052.pdf
Liang, P., Jordan, M. and Klein, D., Learning dependency-based compositional semantics, ACL 2011
aclweb.org/anthology-new/P/P11/P11-1060.pdf
Manshadi, M. and Allen, J. Unrestricted quantifier scope disambiguation, ACL Textgraphs Wkshp, 2011
aclweb.org/anthology-new/W/W11/W11-1108.pdf

Optional More Background

Bos, J & P. Blackburn, Working with Discourse Representation Theory, 2004
homepages.inf.ed.ac.uk/jbos/comsem/book2.html
Turney P. & P. Pantel, From frequency to meaning: vector space models of semantics, JAIR, 37, 141–188, 2010
arxiv.org/pdf/1003/1141
Richardson, M. & P. Domingos, Markov logic networks, ML 62, 107–136, 2006
citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.170.7952.pdf