Formal Semantics in the Real World

J. Bos

Linguistic Computing Laboratory
Department of Computer Science
University of Rome “La Sapienza”, Italy
bos@di.uniroma1.it

Formal methods for the analysis of the meaning of natural language expressions have long been restricted to the ivory tower built by semanticists, logicians, and philosophers of language. It is only in exceptional cases that these methods make their way straight into open-domain natural language processing tools. Recently, however, this situation has changed. Thanks to (i) the development of treebanks, i.e., large collections of texts annotated with syntactic structures, (ii) robust statistical parsers trained on such treebanks, and (iii) the development of large-scale semantic lexica such as WordNet [1], VerbNet [2], PropBank [3], and FrameNet [4], we now have witnessed the development of wide-coverage systems that are able to produce formal semantic representations for open-domain texts.

One such system, developed by myself over the last four years, is Boxer, which follows the principles of Discourse Representation Theory (DRT) to construct and represent meaning of natural languages texts [5,6]. DRT is a formal theory of meaning, initially proposed by Hans Kamp [7] to solve various problems related to anaphoric pronouns. Throughout the years DRT faced various extensions and improvements and by now covers a wide range of semantic issues including plurals and tense [8], discourse segmentation and rhetorical structure [9], and presupposition [10]. Boxer constructs Discourse Representation Structures (DRSs, which are graphically displayed as boxes) with the help of Combinatory Categorial Grammar (CCG) for producing syntactic structure [11] and a typed lambda calculus to specify the syntax-semantics interface [12]. In conjunction with a robust CCG parser [13,14], Boxer achieves very high coverage (> 98% on the Wall Street Journal sections of the Penn Treebank [15]) on newswire text producing DRSs with neo-Davidsonian predicate-argument structure. These DRSs can be translated into standard first-order logic syntax and then fed into automated theorem provers and model builders to check for logical consistency or informativeness [12,16].

The existence of systems like Boxer is clear evidence that practicing formal semantics is not bound to pencil and paper exercises anymore, nor to implementations covering relatively small fragments of natural language. A case in point is the use of Boxer in real-world applications such as open-domain question answering [17]. These developments mark a milestone in the development of computational linguistics in general and computational semantics in particular. They also trigger new research questions, directions and challenges, including the identification of gaps between theory and practice, the inclusion of background
knowledge, transferring theoretical ideas developed in isolation within different logical formalisms into one unifying framework \cite{bos}, and the issue of evaluation.

In particular the evaluation issue is of major importance for further progress in the field. Modelling all nuances of meaning is an immense task — perhaps even impossible. The representations that Boxer produces for a text, as any rival system would, only characterise an approximation of its meaning. An interesting question to ask then is how good this approximation is. How do we measure the semantic adequacy of systems like Boxer that claim are able to compute meaning? A timely question, but despite various proposals aiming to deal with this issue, as yet we cannot answer this question satisfactorily. Comparing a system’s output with gold-standard semantic representations would be an obvious choice but annotated semantic corpora simply don’t exist. Most promising are probably theory-neutral evaluation techniques such as recognising textual inference that we know from the FRACAS project \cite{fracas}, Monz and De Rijke \cite{monz}, and the recent PASCAL challenges \cite{pascal1,pascal2,pascal3}. But such exercises are either considered artificial or fail to isolate semantic competence in systems \cite{artificial}.

References