Open-Domain Semantic Parsing with Boxer

Johan Bos
Center for Language and Cognition
University of Groningen
johan.bos@rug.nl

Abstract

Boxer is a semantic parser for English texts with many input and output possibilities, and various ways to perform meaning analysis based on Discourse Representation Theory. This involves the various ways that meaning representations can be computed, as well as their possible semantic ingredients.

1 Introduction

In this paper I present the capabilities of the open-domain semantic parser Boxer. Boxer is distributed with the C&C text processing tools (Curran et al., 2007), and its main characteristics were first described in my earlier work (Bos, 2008). The roots of the current version of Boxer go back even further, long before Boxer was officially released to the community (Bos et al., 2004; Bos, 2001).

Boxer distinguishes itself from other semantic parsers in that it produces formal meaning representations (compatible with first-order logic) while reaching wide coverage, and is therefore used in a range of applications (Basile et al., 2012; Bjerva et al., 2014). To get an idea of what Boxer does, consider the input and output in Figure 1.

```
John did not go to school .
 ------------------
|x1                |
|..................|
|named(x1,john,per)|
|     ------------ |
|    |e2 x3       ||
| ¬  |............||
|    |go(e2)      ||
|    |agent(e2,x1)||
|    |school(x3)  ||
|    |to(e2,x3)   ||
|    |------------||
|------------------|
```

Figure 1: Example of Boxer's input and output.

Here, the input is a simple sentence, and Boxer's output a formal interpretation of this sentence: there is a person x_1 named "john", and it is not the case that there is a school-going-event e_2 that involves the entities x_1 (John) and a school (denoted by entity x_3). But Boxer has a lot more to offer, and what this paper contributes (and adds with respect to previous publications) is a fine-grained description of the many possibilities that Boxer provides for the formal semantic analysis of text processing.

2 Interface Formats

The input of the Boxer system is a syntactic analysis in the form of a derivation of combinatorial categorial grammar, CCG (Steedman, 2001). This input can be augmented in order to incorporate information of external language technology components. The output is a meaning representation, produced in a variety of standard formats.

2.1 Input

Boxer requires a syntactic analysis of the text in the form of CCG-derivations, every sentence corresponding to one CCG derivation. The derivation itself is represented as a `ccg/2` Prolog term, comprising a sentence identifier and a recursively built structure of combinatorial rules (such as `fa/3`, `ba/3`, and so on), and terminals (the lexical items). All combinatorial rules of CCG are supported, including the generalized composition rules and the type-changing rules introduced in CCGbank (Hockenmaier and Steedman, 2007).

The terminals are captured by a Prolog term consisting of the CCG category (Boxer implements about 600 different lexical category types), the token, its lemma, and part-of-speech. Information of external tools can also be included here, such as word sense disambiguation, thematic role labelling, noun-noun compound interpretation, or reference resolution.

```
sem(1,[1001:[tok:'John',pos:'NNP',lemma:'John',namex:'I-PER'],
       1002:[tok:did,pos:'VBD',lemma:do,namex:'O'],
       1003:[tok:not,pos:'RB',lemma:not,namex:'O'],
       1004:[tok:go,pos:'VB',lemma:go,namex:'O'],
       1005:[tok:to,pos:'TO',lemma:to,namex:'O'],
       1006:[tok:school,pos:'NN',lemma:school,namex:'O'],
       1007:[tok:'.',pos:'.',lemma:'.',namex:'O']],
    b2:drs([b1:[]:x1],
           [b1:[1001]:named(x1,john,per,nam),
            b2:[1003]:not(b3:drs([b3:[]:e1,b3:[]:x2],
                                 [b3:[1004]:pred(e1,go,v,0),
                                  b3:[]:role(e1,x1,agent,1),
                                  b3:[1006]:pred(x2,school,n,0),
                                  b3:[1005]:rel(e1,x2,to,0)]))])).
```

Figure 2: Boxer's output in Prolog format, for "John does not go to school."

Any parser can be used to support Boxer, as long as it produces CCG derivation in the required Prolog format. The standard parser used in tandem with Boxer is that of the C&C tools (Clark and Curran, 2004). Alternatively, other parsers can be used, such as EasyCCG (Lewis and Steedman, 2014). The lemmas can be provided by off-the-shells tools like morpha (Minnen et al., 2001).

2.2 Output

The standard output is a meaning representation in the form of a Discourse Representation Structure (Kamp and Reyle, 1993). This output is standard shown in Prolog format, but can also be produced in XML (with the `--format xml` option). Output can also be suppressed, with `--format no`, in case only human-readable output is wanted.

For the user's convenience, the meaning can also be displayed in boxed format (with the `--box true` option), as shown above. In combination with `--instantiate true`, this yields convenient names for discourse referents that appear in the boxes. Additionally, with the `--ccg true` option, a pretty-printed version of the input CCG-derivation is presented to the user.

3 Semantic Frameworks

3.1 Semantic Theory

The backbone of Boxer's meaning representations is provided by Discourse Representation Theory, DRT (Kamp and Reyle, 1993). Boxer follows the theory closely (`--theory drt`), except with respect to (i) event semantics, where it adopts a neo-Davidsonian approach, and (ii) the analysis of sentential complements, where Boxer follows an analysis based on modal logic (Bos, 2004).

By default, Boxer produces a meaning representation for every sentence in the input. However, with `--integrate true` it computes a single meaning representation spanning all sentences, with separate boxes corresponding to all sentences. Instead, using `--theory sdrt`, a Segmented Discourse Representation Structure is produced, following SDRT (Asher, 1993).

3.2 Meaning Translations

The meaning representations of Discourse Representation Theory can be shaped in different ways, and Boxer supports several of these possibilities. The standard representations are DRSs (Discourse Representation Structures, the boxes, selected with `--semantics drs`). Alternatively DRSs can be shown as Projective DRSs (Venhuizen et al., 2013) using `--semantics pdrs`, where each DRS is labelled with a pointer, and each DRS-condition receives a pointer to the DRS in which it appears.

For some applications and users with different mind-sets, Boxer comes with an option to translate DRSs into other types of meaning representation. First of all, with `--semantics fol`, Boxer supports the well-known translation from boxes to first-order logic (Kamp and Reyle, 1993; Bos, 2004), or to DRSs in the form of graphs (Basile and Bos, 2013), when invoked with `--semantics drg`. Secondly, the meaning representations can be translated into flat logical forms, as proposed in Jerry Hobbs's framework (Hobbs, 1991), with `--semantics tacitus`. Note that not all of these translations are necessary meaning-preserving, becauce of the differences in expressive power between the formalisms.

4 Meaning Details

The devil is in the detail. Indeed, to get the most out of Boxer, it is important to know what features it offers to compute meaning representations.

4.1 Linguistic Features

Copula Notorious among computational semanticists is the analysis of the copula. Boxer gives two options: to interpret the copula as were it an ordinary transitive verb (`--copula false`), or by introducing an equality symbol between two entities (`--copula true`). The latter option has as advantage that certain inferences can directly be drawn, but as disadvantage that some nuances of meaning are lost (i.e., the distinction between *John is a teacher* and *John was a teacher*).

Multiword Expressions Boxer provides two ways to represent compound proper names. With `--mwe no` a compound name such as *Barack Obama* is represented by two naming conditions (with the non-logical symbols `barack` and `obama`), and with `--mwe yes` as a single naming condition (with symbol `barack~obama`).

Noun–Noun Compounds Noun–noun compounds are interpreted as two entities that form a certain relation. By default, Boxer picks the generic prepositional *of*–relation. With `--nn true`, Boxer attempts to disambiguate noun-noun compound relations by selecting from a set of prepositional relations (Bos and Nissim, 2015). For instance, *beach house* would be interpreted as: house(x) $\wedge$ beach(y) $\wedge$ at(x,y).

Reference Resolution By default Boxer doesn't resolve pronouns or other referential expressions, but with `--resolve true`, Boxer attempts to resolve pronouns, proper names and definite descriptions (currently using a rule-based approach). The foundational algorithm to accomplish this is based on Van der Sandt's theory of presupposition projection (Van der Sandt, 1992). The discourse referents of the selected antecedents are unified with those of the referential expression.

Thematic Role Labelling As mentioned above, Boxer follows a neo-Davidsonian approach to event semantics. This means that events (usually triggered by verbs) introduce discourse referents, and these are related to discourse referents of participants by two-place relations, the thematic roles. Standard (`--roles proto`) these roles are picked from a set of five proto-roles: agent, theme, topic, recipient, and experiencer. A more fine-grained inventory of roles is employed with `--roles verbnet`, producing thematic roles as provided by VerbNet (Kipper et al., 2008). This is done by mapping the obtained proto-roles to VerbNet roles, using a simple deterministic approach in the semantic lexicon of Boxer.

4.2 Logical Features

Eliminating Equality In some cases equality symbols can be elimated from the meaning representation, resulting in a logically equivalent logical form. This is possible, for instance, when the two variables within an equality relation are bound by discourse referents introduced in the same DRS as the equality condition. Equality conditions are introduced by a range of lexical entries, but in the final meaning representation they don't play a fundamental role. With `--elimeq true` such equality conditions are removed and their corresponding discourse referents unified.

Modal Modal expressions (as introduced by modal adverbs or modal verbs) can be made explicit in the meaning representation by invoking `--modal true`. This triggers two additional complex DRS-conditions formed by the unary box and diamond operators from modal logic, expressing necessity (univerally quantifying over possible worlds) and possibility (existentially quantifying over possible worlds). This option also has an effect on the translation to first-order logic, and when used in combination with `--semantics fol` the translation to modal first-order logic is used (with reification over possible worlds).

Tense The standard reference textbook for Discourse Representation Theory has an extensive analysis of various tenses found in the English language (Kamp and Reyle, 1993). Boxer aims to reproduce this analysis with `--tense true`. This involves additional relations and discourse referents related to the events introduced by the text.

5 Conclusion

I have outlined a large set of possibilities that the semantic parser Boxer offers. These concern input and output modalities, as well as the level of detail of meaning interpretation. I will demonstrate a selection of these features at the conference.

References

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.

Valerio Basile and Johan Bos. 2013. Aligning formal meaning representations with surface strings for wide-coverage text generation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 1–9, Sofia, Bulgaria.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Negation detection with discourse representation structures. In *The First Joint Conference on Lexical and Computational Semantics (*SEM 2012 Shared Task)*, pages 301–309, Montreal, Canada.

Johannes Bjerva, Johan Bos, Rob Van der Goot, and Malvina Nissim. 2014. The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 642–646, Dublin, Ireland.

Johan Bos and Malvina Nissim. 2015. Uncovering noun-noun compound relations by gamification. In Beáta Megyesi, editor, *Proceedings of the 20th Nordic Conference of Computational Linguistics*.

Johan Bos, Stephen Clark, Mark Steedman, James R. Curran, and Julia Hockenmaier. 2004. Wide-Coverage Semantic Representations from a CCG Parser. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*, pages 1240–1246, Geneva.

Johan Bos. 2001. DORIS 2001: Underspecification, Resolution and Inference for Discourse Representation Structures. In Patrick Blackburn and Michael Kohlhase, editors, *ICoS-3, Inference in Computational Semantics*, pages 117–124.

Johan Bos. 2004. Computational Semantics in Discourse: Underspecification, Resolution, and Inference. *Journal of Logic, Language and Information*, 13(2):139–157.

Johan Bos. 2008. Wide-Coverage Semantic Analysis with Boxer. In J. Bos and R. Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 277–286. College Publications.

Stephen Clark and James R. Curran. 2004. Parsing the WSJ using CCG and Log-Linear Models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL '04)*, pages 104–111, Barcelona, Spain.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic.

Jerry R. Hobbs. 1991. SRI international's TACITUS system: MUC-3 test results and analysis. In *Proceedings of the 3rd Conference on Message Understanding, MUC 1991, San Diego, California, USA, May 21-23, 1991*, pages 105–107.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: a corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT*. Kluwer, Dordrecht.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40.

Mike Lewis and Mark Steedman. 2014. A* ccg parsing with a supertag-factored model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 990–1000, Doha, Qatar, October. Association for Computational Linguistics.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of english. *Journal of Natural Language Engineering*, 7(3):207–223.

Mark Steedman. 2001. *The Syntactic Process*. The MIT Press.

Rob A. Van der Sandt. 1992. Presupposition Projection as Anaphora Resolution. *Journal of Semantics*, 9:333–377.

Noortje Venhuizen, Johan Bos, and Harm Brouwer. 2013. Parsimonious semantic representations with projection pointers. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 252–263, Potsdam, Germany.