

Wide-Coverage Semantic Analysis with Boxer

Johan Bos

University of Rome “La Sapienza” (Italy)

email: bos@di.uniroma1.it

Abstract

Boxer is an open-domain software component for semantic analysis of text, based on Combinatory Categorical Grammar (CCG) and Discourse Representation Theory (DRT). Used together with the C&C tools, Boxer reaches more than 95% coverage on newswire texts. The semantic representations produced by Boxer, known as Discourse Representation Structures (DRSs), incorporate a neo-Davidsonian representations for events, using the VerbNet inventory of thematic roles. The resulting DRSs can be translated to ordinary first-order logic formulas and be processed by standard theorem provers for first-order logic. Boxer's performance on the shared task for comparing semantic representations was promising. It was able to produce complete DRSs for all seven texts. Manually inspecting the output revealed that: (a) the computed predicate argument structure was generally of high quality, in particular dealing with hard constructions involving control or coordination; (b) discourse structure triggered by conditionals, negation or discourse adverbs was overall correctly computed; (c) some measure and time expressions are correctly analysed, others aren't; (d) several shallow analyses are given for lexical phrases that require deep analysis; (e) bridging references and pronouns are not resolved in most cases. Boxer is distributed with the C&C tools and freely available for research purposes.

1 Introduction

Boxer is an open-domain tool for computing and reasoning with semantic representations. Based on Discourse Representation Theory (Kamp and Reyle, 1993), Boxer is able to construct Discourse Representation Structures (DRSs for short, informally called “boxes” because of the way they are graphically displayed) for English sentences and texts. There is a translation from DRSs to first-order formulas, which opens the way to perform inference by including automated reasoning tools such as theorem provers and model builders (Blackburn and Bos, 2005).

2 Theory

2.1 Combinatory Categorical Grammar

As a preliminary to semantics, we need syntax. Boxer implements a syntax-semantics interface based on Combinatory Categorical Grammar, CCG (Steedman, 2001). CCG lends itself extremely well for this task because it is lexically driven and has only few “grammar” rules, and not less because of its type-transparency principle, which says that each syntactic type (a CCG category) corresponds to a unique semantic type (a lambda-expression). Because the syntax-semantics is clearly defined, the choice of logical form can be independent of the categorial framework underlying it. Steedman uses simple predicate argument structures expressed via the untyped lambda calculus to illustrate the construction of logical forms in CCG (Steedman, 2001). We instead opt for Discourse Representation Theory, a widely accepted sophisticated formal theory of natural language meaning dealing with a large variety of semantic phenomena.

2.2 Discourse Representation Theory

DRT is a formal semantic theory originally designed by Kamp to cope with anaphoric pronouns and temporal relations (Kamp, 1981). DRT uses an explicit intermediate semantic representation, called DRS (Discourse Representation Structure), for dealing with anaphoric or other contextually sensitive linguistic phenomena such as ellipsis and presupposition. We choose DRT because it has established itself as a well-documented formal theory of meaning, covering a number of semantic phenomena ranging from pronouns, abstract anaphora, presupposition, tense and aspect, propositional attitudes, to plurals (Kamp and Reyle, 1993; Asher, 1993; Van der Sandt, 1992).

In terms of expressive power, three different kinds of representations are distinguished in Boxer:

1. Discourse Representation Structures (DRSs)
2. Underspecified DRSs (DRSs + merge + alfa)
3. λ -DRSs (UDRSs + lambda + application)

DRSs are the representations corresponding to natural language sentences or texts. This is the core DRT language compatible with first-order logic. The DRS language employed by Boxer is a subset of the one found in Kamp and Reyle (1993). We define the syntax of DRSs below with the help of Backus-Naur form, where non-terminal symbols are enclosed in angle brackets. The non-terminal $\langle \text{ref} \rangle$ denotes a discourse referent, and $\langle \text{sym}_n \rangle$ an n -place predicate symbol.

```

<expe> ::= <ref>
<expt> ::= <drs>
<drs> ::= 

|              |
|--------------|
| <ref>*       |
| <condition>* |


<condition> ::= <basic> |
               <complex>
<basic> ::= <sym1>(<expe>) |
            <sym2>(<expe>, <expe>) |
            <named>(<expe>, <nam>, <sort>)
<complex> ::= ¬<expt> |
             <expt>⇒<expt> |
             <expt>∨<expt> |
             <ref>:<expt>

```

DRSs are structures comprising two parts: 1) a set of discourse referents; and 2) a set of conditions constraining the interpretation of the discourse referents. Conditions can be simple properties of discourse referents, express relations between them, or be complex, introducing (recursively) subordinated DRSs.

The standard version of DRT formulated in Kamp & Reyle incorporates a Davidsonian event semantics (Kamp and Reyle, 1993), where discourse referents can also stand for events and be referred to by anaphoric expressions or constrained by temporal relations. The neo-Davidsonian system, as implemented in Boxer, uses the inventory of roles proposed by VerbNet (Kipper et al., 2008), and has some attractive formal properties (Dowty, 1989). There is only one way to state that an individual is participating in an event—namely by relating it to the event using a binary relation expressing some thematic role. Furthermore, the approach clearly distinguishes the participants of an event by the semantic roles they bear. Finally, it also allows us to characterize the meaning of thematic roles independently of the meaning of the verb that describes the event.

We won't show the standard translation from DRS to FOL here (Blackburn et al., 2001; Bos, 2004; Kamp and Reyle, 1993). Intuitively, translating DRSs into first-order formulas proceeds as follows: each discourse referent is translated as a first-order quantifier, and all DRS-conditions are translated into a conjunctive formula of FOL. Discourse referents usually are translated to existential quantifiers, with the exception of those declared in antecedents of implicational DRS-conditions, that are translated as universal quantifiers. Obviously, negated DRSs are translated as negated formulas, disjunctive DRSs as disjunctive formulas, and implicational DRSs as formulas with material implication.

Boxer outputs either resolved semantic representations (in other words, completely disambiguated DRSs), or underspecified representations, where some ambiguities are left unresolved in the semantic representation. This level of representation is referred to as underspecified DRS, or UDRS for short. It is a small extension of the DRS language given in the previous section and is defined as follows:

```

<expt> ::= <udrs>
<udrs> ::= <drs> | (<expt>;<expt>) | (<expt>α<expt>)

```

Note here that expressions of type *t* are redefined as UDRSs. UDRSs are either ordinary DRSs, DRSs conjoined by the merge (for which we use the semicolon), or

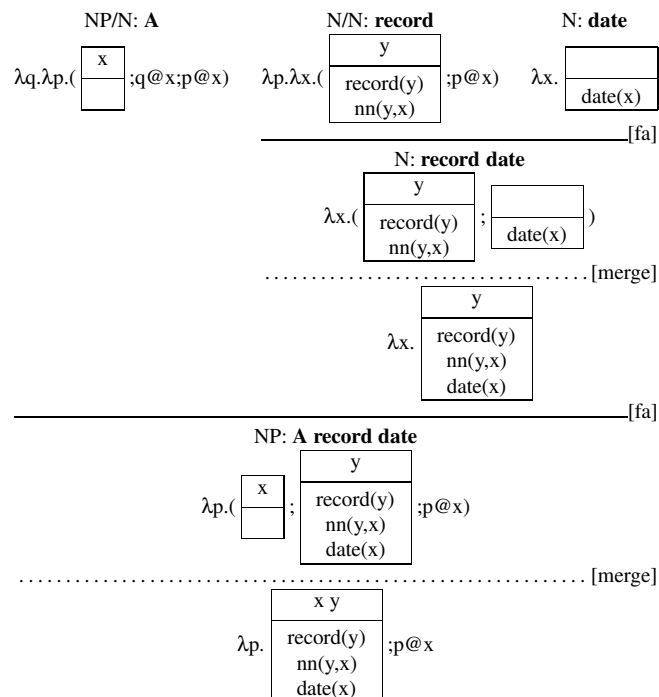


Figure 1: Derivation with λ -DRSs, including β -conversion, for “A record date”. Combinatory rules are indicated by solid lines, semantic rules by dotted lines.

DRS composed by the α -operator. The merge conjoins two DRSs into a larger DRS — semantically the merge is interpreted as (dynamic) logical conjunction. Merge-reduction is the process of eliminating the merge operation by forming a new DRS resulting from the union of the domains and conditions of the argument DRSs of a merge, respectively (obeying certain constraints). Figure 1 illustrates the syntax-semantics interface (and merge-reduction) for a derivation of a simple noun phrase.

Boxer adopts Van der Sandt’s view as presupposition as anaphora (Van der Sandt, 1992), in which presuppositional expressions are either resolved to previously established discourse entities or accommodated on a suitable level of discourse. Van der Sandt’s proposal is cast in DRT, and therefore relatively easy to integrate in Boxer’s semantic formalism. The α -operator indicates information that has to be resolved in the context, and is lexically introduced by anaphoric or presuppositional expressions. A DRS constructed with α resembles the proto-DRS of Van der Sandt’s theory of presupposition (Van der Sandt, 1992) although they are syntactically defined in a slightly different way to overcome problems with free and bound variables, following Bos (2003). Note that the difference between anaphora and presupposition collapses in Van der Sandt’s theory.

The types are the ingredients of a typed lambda calculus that is employed to construct DRSs in a bottom-up fashion, compositional way. The language of lambda-

DRSs is an extension of the language of (U)DRS defined before:

$$\begin{aligned} \langle \text{exp}_e \rangle & ::= \langle \text{ref} \rangle \mid \langle \text{var}_e \rangle \\ \langle \text{exp}_t \rangle & ::= \langle \text{udrs} \rangle \mid \langle \text{var}_t \rangle \\ \langle \text{exp}_\alpha \rangle & ::= (\langle \text{exp}_{(\beta,\alpha)} \rangle @ \langle \text{var}_\beta \rangle) \mid \langle \text{var}_\alpha \rangle \\ \langle \text{exp}_{(\alpha,\beta)} \rangle & ::= \lambda \langle \text{var}_\alpha \rangle . \langle \text{exp}_\beta \rangle \mid \langle \text{var}_{(\alpha,\beta)} \rangle \end{aligned}$$

Hence we define discourse referents as expressions of type e , and DRSs as expressions of type t . We use @ to indicate function application, and the λ -operator to bind free variables over which we wish to abstract.

3 Practice

3.1 Preprocessing

The input text needs to be tokenised with one sentence per line. In the context of this paper, Boxer was put into action after using a combined processing pipeline of the C&C tools consisting of POS-tagging, named entity recognition, and parsing (Curran et al., 2007). The POS tags are used to specify the lexical semantics for ambiguous CCG categories (see below); the named entity tags are transferred to the level of DRSs as well and added as sorts to named discourse referents. An example of a CCG derivation is shown in Figure 2.

```

Cervical cancer          caused          a          virus
---[lex] ---[lex]          by          ---[lex] ---[lex]
n/n      n              is          s:pss\np  (s:pss\np)\(s:pss\np)  np:nb/n n
-----[fa] -----[lex] -----[lex] -----[fa]
n              (s:dcl\np)/(s:pss\np) s:pss\np
-----[tc] -----[lex] -----[ba]
np              s:dcl\np
-----[fa]
s:dcl

```

Figure 2: CCG derivation as generated by the C&C tools

3.2 Lexicon

In CCG, the syntactic lexicon comprises the set of lexical categories. CCGbank hosts more than a thousand different categories. The semantic lexicon defines a suitable mapping from categories to semantic representations. In the context of Boxer, these semantic representations are defined in the shape of lambda-DRSs. Boxer implements almost all categories employed by the C&C parser, which is a subset of the ones found in CCGbank, leaving out extremely rare cases for the sake of efficiency.

Defining the lexical semantics cannot always be done solely on the basis of the category, for one lexical category could give rise to several different semantic interpretations. So we need to take other resources into account, such as the assigned part of speech (PoS), and sometimes the wordform or named entity type associated with the category. For the majority of categories, in particular those that correspond to

open-class lexical items, we also need access to the morphological root of the word that triggered the lexical category.

Although there is a one-to-one mapping between the CCG categories and semantic types — and this must be the case to ensure the semantic composition process proceeds without type clashes — the actual instantiations of a semantic type can differ even within the scope of a single CCG category. For example, the category n/n can correspond to an adjective, a cardinal expression, or even common nouns and proper names (in the compound expressions). In the latter two cases the lexical entry introduces a new discourse referent, in the former two it does not. To account for this difference we also need to look at the part of speech that is assigned to a token.

3.3 Resolution

Boxer implements various presupposition triggers introduced by noun phrases, including personal pronouns, possessive pronouns, reflexive pronouns, emphasising pronouns, demonstrative pronouns, proper names, other-anaphora, definite descriptions. In addition, some aspects of tense are implemented as presupposition triggers, too.

Anaphora and presupposition resolution takes place in a separate stage after building up the representation, following the resolution algorithm outlined in Bos (2003). The current implementation of Boxer aims at high precision in resolution: personal pronouns are only attempted to be resolved to named entities, definite descriptions and proper names are only linked to previous discourse referents if there is overlap in the DRS-conditions of the antecedent DRS and alpha-DRS. If no suitable antecedent can be found, global accommodation of the anaphoric discourse referent and conditions will take place.

Because Boxer has the option to output unresolved DRSs too, it is possible to include external anaphora or coreference resolution components.

3.4 Example Analysis

We illustrate the capabilities of Boxer with the following example text shown below (aka as Text 2 of the shared task).¹ The text consists of three sentences, the second being a coordinated sentence. It contains a passive construction, three pronouns, relative clauses, control verbs, and a presupposition trigger *other*.

Text 2

Cervical cancer is caused by a virus. That has been known for some time and it has led to a vaccine that seems to prevent it. Researchers have been looking for other cancers that may be caused by viruses.

The output of Boxer for this text is shown in Figure 3. Only the box format is shown here — Boxer is also able to output the DRSs in Prolog or XML encodings. It was run without analysing tense and aspect and without discourse segmentation (both of these are possible in Boxer, but still undergo development, and are therefore disregarded here).

As we can see from the example and Boxer's analysis various things go right and various things go wrong. Boxer deals fine with the passive construction (assigned the

¹This text was taken from the Economist Volume 387 Number 8582, page 92. The third sentence has been simplified.

appropriate semantic role), the relative clauses, and the control construction (*vaccine* is the agent of the *prevent* event). It also handles the presupposition trigger anaphorically linking the mention of *other cancers* in the third sentence with the phrase *cervical cancer* in the first sentence, and asserting an inequality condition in the DRS.

Boxer failed to resolve three pronouns correctly. These are all accommodated at the global level of DRS, which is the DRS on the left-hand side in Figure 3. All of the pronouns have textual antecedents: the abstract pronoun *that* in the second sentence refers to the fact declared in the first sentence. The first occurrence of *it* in the second sentence also seems to refer to this fact — the second occurrence of *it* refers to cervical cancer mentioned in the first sentence.

```
bin/boxer --input working/step/text2.ccg --semantics drs --box --resolve --roles verbnnet --format no
%%
%% | x0 x1 x2 | | x3 x4 x5 | | x6 x7 | | x8 x9 x10 x11 | | x13 x14 x15 x16 x17 |
%% |-----| |-----| |-----| |-----| |-----|
%% ( ( thing(x0) | + ( cancer(x3) | + ( know(x6) | + ( lead(x8) | + ( researcher(x13) | )))
%% | neuter(x1) | | cervical(x3) | | time(x7) | | vaccine(x9) | | look(x14) |
%% | neuter(x2) | | cause(x4) | | event(x6) | | seem(x10) | | agent(x14,x13) |
%% | | | | virus(x5) | | theme(x6,x0) | | proposition(x11) | | cancer(x15) |
%% | | | | event(x4) | | for(x6,x7) | | event(x10) | | |
%% | | | | theme(x4,x3) | | | | event(x8) | | |
%% | | | | by(x4,x5) | | | | agent(x8,x1) | | |
%% | | | | | | | | agent(x10,x9) | | | x15 = x3 |
%% | | | | | | | | theme(x10,x11) | | | |
%% | | | | | | | | to(x8,x9) | | | cause(x16) |
%% | | | | | | | | | | virus(x17) |
%% | | | | | | | | | | event(x16) |
%% | | | | | | | | | | theme(x16,x15) |
%% | | | | | | | | | | prevent(x12) |
%% | | | | | | | | | | event(x12) |
%% | | | | | | | | | | agent(x12,x9) |
%% | | | | | | | | | | theme(x12,x2) |
%% | | | | | | | | | | event(x14) |
%%
%%
Attempted: 3. Completed: 3 (100.00%).
```

Figure 3: Boxer output for Shared Task Text 2

4 Performance on Shared Task

Here we discuss the output of Boxer on the Shared Task Texts (Bos, 2008). Boxer was able to produce semantic representation for all text without any further modifications to the software. For each text we briefly say what was good and bad about Boxer’s analysis. (We won’t comment on the performance on the second text, as this is the text proposed by ourselves and already discussed in the previous section.)

Text 1: An object is thrown with a horizontal speed ...

Good: The resulting predicate argument structure was fine overall, including a difficult control construction (“how long does it take the object to fall ...”). The definite description “the object” was correctly resolved. The conditional got correctly analysed.

Bad: The measure phrase “125 m high” got mis-interpreted as noun-noun compound. The definite description “the fall” was not linked to the falling event mentioned before.

Comments: Because there were two questions in this text we parsed it using the C&C parser with the model trained on questions.

Text 3: John went into a restaurant ...

Good: The pronouns were correctly resolved to the proper name “John” rather than “the waiter”, even though this is based on the simple strategy in Boxer to link third-person pronouns to named entities of type human. The coordination construction “warm and friendly” got correctly analysed (distributively), and the control construction “began to read his book” received a proper predicate argument structure.

Bad: Boxer doesn’t deal with bridging references introduced by relational nouns, so expressions like “the corner” were not linked to other discourse entities.

Text 4: The first school for the training of leader dogs ...

Good: The named entities were correctly recognised and classified (locations and proper names). The VP coordination in the first and later sentences was correctly analysed. The expression “this school” got correctly linked to the school mentioned earlier in the text. The time expression “1999” got the right interpretation.

Bad: The adjectives/determiners “first” and “several” didn’t receive a deep analysis. The complex NP “Joao Pedro Fonseca and Marta Gomes” was distributively interpreted, rather than collective. The pronoun “they” wasn’t resolved. The preposition “In” starting the second sentence was incorrectly analysed by the parser.

Text 5: As the 3 guns of Turret 2 were being loaded ...

Good: The discourse structures invoked by the sentence initial adverbs “As” and “When” was correctly computed. Predicate argument structure overall good, including treatment of the relative clauses. The expression “the propellant” was correctly resolved. Time expressions in the one but last sentence got a correct analysis.

Bad: The name “Turret 2” was incorrectly analysed (not as a compound). The adverbs “yet” and “then” got a shallow analysis. The first-person pronoun “I” was not resolved to the crewman.

Comments: The quotes were removed in the tokenisation phase, because the C&C parser, being trained on a corpus without quotes, performs badly on texts containing quotes.

Text 6: Amid the tightly packed row houses of North Philadelphia ...

Good: The named entities were correctly recognised and classified as locations. The various cases of VP coordination all got properly analysed. The numerical and date expressions got correct representations.

Bad: The occurrences of the third-person neuter pronouns were not resolved. The preposition “Amid” was not correctly analysed.

Text 7: Modern development of wind-energy technology and applications ...

Good: Correct interpretation of time expressions “1930s” and “1970s”. Correct predicate argument structure overall.

Bad: “Modern” was recognised as a proper name. The noun phrase “wind-energy technology and applications” was distributively analysed with “wind-energy” only applying to “technology”. The sentence-initial adverb “Since” did not introduce proper discourse structure. The units of measurement in the last two sentences were not

recognised as such. The tricky time expression “mid-80’s” only got a shallow interpretation.

5 Conclusion

Boxer is a wide-coverage system for semantic interpretation. It takes as input a CCG derivation of a natural language expression, and produces formally interpretable semantic representations: either in the form of DRSs, or as formulas of first-order logic. The existence of CCGbank (Hockenmaier, 2003) and robust parsers trained on it (Clark and Curran, 2004; Bos et al., 2004) make Boxer a state-of-the-art open-domain tool for deep semantic analysis.

Boxer’s performance on the shared task for comparing semantic representations was promising. It was able to produce DRSs for all texts. We can’t quantify the quality of Boxer’s output, as we don’t have gold standard representations at our disposal. Manually inspecting the output gives us the following impression:

- computed predicate argument structure is generally of good quality, including hard constructions involving control or coordination;
- discourse structure triggered by conditionals, negation or discourse adverbs is overall correctly computed;
- some measure and time expressions are correctly analysed, others aren’t;
- several shallow analyses are given for lexical phrases that require deep analysis;
- bridging references and pronouns are not resolved in most cases; but when they are, they are mostly correctly resolved (high precision at the cost of recall).

Finally, a comment on availability of Boxer. All sources of Boxer are available for download and free of non-commercial use. It is distributed with the C&C tools for natural language processing (Curran et al., 2007), which are hosted on this site:

<http://svn.ask.it.usyd.edu.au/trac/candc/wiki/boxer>

References

- Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer Academic Publishers.
- Blackburn, P. and J. Bos (2005). *Representation and Inference for Natural Language. A First Course in Computational Semantics*. CSLI.
- Blackburn, P., J. Bos, M. Kohlhase, and H. de Nivelle (2001). Inference and Computational Semantics. In H. Bunt, R. Muskens, and E. Thijsse (Eds.), *Computing Meaning Vol.2*, pp. 11–28. Kluwer.
- Bos, J. (2003). Implementing the Binding and Accommodation Theory for Anaphora Resolution and Presupposition Projection. *Computational Linguistics* 29(2), 179–210.

- Bos, J. (2004). Computational Semantics in Discourse: Underspecification, Resolution, and Inference. *Journal of Logic, Language and Information* 13(2), 139–157.
- Bos, J. (2008). Introduction to the Shared Task on Comparing Semantic Representations. In J. Bos and R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, pp. 257–261. College Publications.
- Bos, J., S. Clark, M. Steedman, J. R. Curran, and J. Hockenmaier (2004). Wide-Coverage Semantic Representations from a CCG Parser. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*, Geneva.
- Clark, S. and J. 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)*, Barcelona, Spain.
- Curran, J., S. Clark, and J. Bos (2007, June). 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*, Prague, Czech Republic, pp. 33–36. Association for Computational Linguistics.
- Dowty, D. (1989). On the semantic content of the notion thematic role. In *Properties, Types, and Meanings*, Volume 2. Kluwer.
- Hockenmaier, J. (2003). *Data and Models for Statistical Parsing with Combinatory Categorical Grammar*. Ph. D. thesis, University of Edinburgh.
- Kamp, H. (1981). A Theory of Truth and Semantic Representation. In J. Groenendijk, T. M. Janssen, and M. Stokhof (Eds.), *Formal Methods in the Study of Language*, pp. 277–322. Amsterdam: Mathematical Centre, Amsterdam.
- Kamp, H. and U. Reyle (1993). *From Discourse to Logic; An Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and DRT*. Dordrecht: Kluwer.
- Kipper, K., A. Korhonen, N. Ryant, and M. Palmer (2008). A large-scale classification of english verbs. *Language Resources and Evaluation* 42(1), 21–40.
- Steedman, M. (2001). *The Syntactic Process*. The MIT Press.
- Van der Sandt, R. (1992). Presupposition Projection as Anaphora Resolution. *Journal of Semantics* 9, 333–377.