The Meaning Factory at SemEval-2016 Task 8: Producing AMRs with Boxer

Johannes Bjerva
CLCG
University of Groningen
j.bjerva@rug.nl

Johan Bos
CLCG
University of Groningen
johan.bos@rug.nl

Hessel Haagsma
CLCG
University of Groningen
hessel.haagsma@rug.nl

Abstract

We participated in the shared task on meaning representation parsing (Task 8 at SemEval-2016) with the aim of investigating whether we could use Boxer, an existing open-domain semantic parser, for this task. However, the meaning representations produced by Boxer, Discourse Representation Structures, are considerably different from Abstract Meaning Representations, AMRs, the target meaning representations of the shared task. Our hybrid conversion method (involving lexical adaptation as well as post-processing of the output) failed to produce state-of-the-art results. Nonetheless, F-scores of 53% on development and 47% on test data (50% unofficially) were obtained.

1 Introduction

With the currently increasing interest in semantic parsing, and the diversity of the meaning representations being used, an important challenge is to adapt existing semantic parsers for different semantic representations. Shared Task 8 of the SemEval-2016 campaign for semantic evaluation is an interesting venue for this, where a system is given an English sentence and has to produce an Abstract Meaning Representation (AMR) for it.

We participated in this shared task with a system rooted in formal semantics based on Discourse Representation Theory (DRT). In particular, we were interested in finding out whether the representations from DRT (Kamp, 1984; Kamp and Reyle, 1993), Discourse Representation Structures (DRSs), could be easily converted into AMRs. In this paper we outline our method, which is based on the semantic parser Boxer (Bos, 2008; Bos, 2015), and then present and discuss our results.

2 Background

Before we outline our method, we will say a little about the open-domain semantic parser that we used in this shared task. We also give an overview of the differences between the meaning representations produced by Boxer and those that are required for the shared task. To get a first taste of these differences, compare the analysis of *'All equipment will be completely manufactured'* carried out by Boxer (Figure 1) and that of the gold-standard AMR (Figure 2).

```
 ___________________________________
|                                   |
|...................................|
|  _____________   _______________  |
| |x1           | |e1 s1          | |
| |.............| |...............| |
| |equipment(x1)|>|manufacture(e1)| |
| |_____________| |Manner(e1,s1)  | |
|                 |Theme(e1,x1)   | |
|                 |complete(s1)   | |
|                 |_______________| |
|___________________________________|
```

Figure 1: DRS, as produced by Boxer.

```
(m / manufacture-01
   :ARG1 (e2 / equipment
            :mod (a2 / all))
   :ARG1-of (c / complete-02))
```

Figure 2: Gold-standard AMR.

Proceedings of SemEval-2016, pages 1179–1184,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

2.1 Boxer

The semantic parser that we employed is Boxer (Bos, 2008; Bos, 2015). It is the last component in the C&C tools pipeline (Curran et al., 2007), comprising a tokenizer (Evang et al., 2013), POS-tagger, lemmatizer (Minnen et al., 2001), and a robust parser for CCG, Combinatory Categorial Grammar (Steedman, 2001). Overall, this parsing framework shows many points of contact with the recent work by Artzi et al. (2015), who also use CCG coupled with a formal compositional semantics.

Boxer produces semantic representations based on Discourse Representation Theory (Kamp and Reyle, 1993), known as Discourse Representation Structures (DRSs), as Figure 1 shows. A DRS is a first-order representation, i.e., expressible with first-order logic. Various notations are possible, but widely used are the box-like representations shown in Figure 1. Boxes display scopes of discourse referents and contain properties of and relations between discourse referents. They are recursive structures, hence a box may contain other boxes.

2.2 Abstract Meaning Representations

At first glance, an AMR looks quite different from a DRS. Usually, an AMR is displayed as a directed graph with a unique root (Figure 2). However, it is also possible to view an AMR as a recursive structure, and then DRS and AMR have more in common than one perhaps would initially realize (Bos, 2016).

The variables in an AMR correspond to discourse referents in a DRS. The colon-prefixed symbols in an AMR are similar to the two-place relation symbols in a DRS. And the forward slashes in an AMR correspond to one-place predicates in a DRS. So the main commonalities between a DRS (as produced by Boxer) and an AMR (as used at the SEMEVAL-2016 shared task) are:

- both use a neo-Davidsonian event semantics;
- both are recursive meaning representations;
- both expect normalization of date expressions.

There are also some obvious differences between DRS and AMR. Some of them are theoretical and have to do with the expressive power of the chosen formalism. Others have to do with relatively arbitrary choice of labels and encoding of meanings. The most important differences are:

- AMR has no explicit structures for universal quantification and negation;
- AMR expects different labels for thematic roles (Boxer uses the VerbNet inventory);
- AMR assigns no scope for propositional meanings;
- AMR is strongly event-oriented (verbalization);
- AMR has flat lists of coordinated structures;
- AMR has symbol grounding by wikification (for named entities).

These are substantial differences posing a serious challenge when mapping DRSs to AMRs. In the next section we describe how we go about doing this.

3 Method

3.1 Pre-processing and Tokenisation

All input texts were normalized prior to semantic processing by our pipeline. First of all, double quotes were removed from sentences because they do not contribute to AMR components but they might give rise to suboptimal syntactic parses. Secondly, brackets containing unit conversions were removed (i.e., *30 yards (27 meters)* became *30 yards*) because the converted information does not show in gold-standard AMRs.

Tokenization was done using the Elephant tokenizer (Evang et al., 2013). Some of the documents that were supplied for the shared task had already undergone tokenization, however. Therefore, prior to tokenization, we determine per document whether or not tokenization is needed. This is done with a simple heuristic, applying tokenization if the document does not contain a full stop or comma with whitespace on both sides.

3.2 Lexical anticipation

Our key idea was to map DRSs as output by Boxer to AMR, and doing so in a systematic, principled way. However, during the implementation process it became clear that this mapping would become much easier when certain conversions would have been already made in Boxer's semantic lexicon. An obvious case is determiners, which receive an elaborate analysis in DRS but a minimal treatment in AMR. Anticipating this in the lexicon saves error-prone conversion steps later in the processing pipeline.

Apart from determiners, lexical conversion (i.e., altering the lexical semantics in order to get closer to AMR structures) was carried out for certain punctuation symbols (question and exclamation marks), all cases of coordination, for some non-logical symbols (for instance, contrastive discourse relations and conditionals), there-insertion, personal and possessive pronouns, demonstratives and quantifiers, comparatives and superlatives, certain temporal modifiers, and copula constructions.

3.3 From DRS to AMR

The conversion from DRS to AMR was implemented using a recursive translation function. Apart from some core translation rules mapping DRS to AMR constructs, there is also a set of rules that work on specific phenomena: modal operators in DRSs are mapped to events (recommend-01 and possible-01); the negation operator is mapped to `polarity-`, and disjunction to an `or`-instance with `op1` and `op2` relations.

Boxer's thematic roles are mapped to ARG0 (Actor) ARG1 (Theme, Topic) or ARG2 (recipient). In addition, we took advantage of Ulf Hermjakob's lists of `have-rel-role` and `have-org-role` predicates to rewrite roles when needed. A similar resource was used to cope with deverbalizations.

The outcome of the mapping is an AMR with possibly more than one root. Therefore the conversion also involves inversion of AMR roles until an AMR with a unique root is obtained. This is a non-trivial process and does not always succeed. In such cases only parts of the AMR are produced as output.

3.4 Re-labelling

An additional post-processing step consisted of changing labels where our output AMRs consistently differed from those in the training data. After processing each document in the training data with Boxer, Smatch (v 2.0.2) was used to obtain all matching triples for each AMR parse pair. Using these triples, we calculate counts of (BOXER-RELATION, GOLD-RELATION, COUNT). If Boxer consistently outputs a relation erroneously, we replace all occurrences of that label with the correct label from the training data.

Examples of phenomena that require re-labelling are: intensifiers, locative adjectives, ordinals, temporal adverbs, morphological mappings of symbols (for instance, historical → history), temporal roles, putting names together, time expressions, units of measurement, nationalities, modal adverbs, verbalizations, negation affixes, and abbreviations.

3.5 Wikification

Wikification was done as a post-processing step: each name-relation produced by boxer was initialized with an empty wiki (`:wiki -`). The value of the wiki was acquired by wikifying the whole sentence and then matching the wikification output to the name. One exception to this regards demonyms, which were assigned the correct wikification by Boxer already.

Wikification was done using DBPedia Spotlight (Daiber et al., 2013) through the web service[1], because of its high coverage and ease-of-use. Since we used Spotlight only for wikification, and not for NER, a high recall was more important than a high precision. The confidence parameter was optimized on the development set, and the optimal value turned out to be 0.3. This is a low value, which yields a large number of annotations for each sentence, a large proportion of which are incorrect.

The wikification output was then matched to the names in the sentence by using exact string matching, and if that failed, by matching on prefixes. Performance was high, with accuracy of wikification only, tested with the gold-standard AMRs at around 76%. In terms of AMR-parsing F-scores, wikification yielded gains of 2% to 4% on the development set, depending on the nature of the data and the quality of Boxer's NER. We also experimented with the Illinois Wikifier (Ratinov et al., 2011), but this did not yield any improvements over DBPedia Spotlight.

[1] http://spotlight.sztaki.hu:2222/rest/annotate

Table 1: F-scores on the *test* part of the released training data.

	DFA	Xinhua	Consensus	Bolt	Proxy
Boxer	39.9	**57.2**	45.8	47.0	56.0
JAMR	**47.5**	52.8	**49.6**	**48.7**	**60.2**

4 Results and Discussion

4.1 Overall Results

We obtained an F-score of 47% in the official scoring. Due to an error early in our pipeline script, a large amount of our parsing mistakes were caused by erroneous tokenization. Correcting this bug results in an F-score of 50% on the official evaluation data (calculated with Smatch v2.0.2 using 4 restarts).

Table 1 shows the F-scores we obtained on the *test* portion of the data set released for system development for this task. We compare our system with the JAMR parser, trained following released instructions on the *training* portion of the released data. Although our parser obtains a lower score on most sub-corpora, we are able to outperform the JAMR parser on the Xinhua sub-corpus.

4.2 Error Analysis

An analysis of the mistakes made on the gold test set reveals that some mistakes can be attributed to annotation mistakes. Figure 3 shows an example in which our AMR is arguably better than the gold AMR. In the sentence *'They are thugs and deserve a bullet.'*, the ones deserving a bullet should not be *all thugs* as in the gold parse, but the referent of *they*, as in our output. Figure 4 shows a similar instance, in which *A protester* is incorrectly assigned the modifier `:quant 1` in the gold parse. (An anonymous reviewer of an earlier version of this article noted that "the AMR gold seems correct, even though I would have probably accepted the Boxer output as correct, at least without knowing more about the pragmatic context of the sentence.". We don't agree here, as many similar cases in the corpus are not annotated with the same attribute, and since there is no context, it is impossible to infer the less likely specific-indefinite reading.)

Another portion of the mistakes made by our system can be attributed to wrong choices of senses, arguments and coordination mistakes. Figure 5 shows an example in which we make a coordination mistake with the noun-noun compound *security force*, and interpret this as a possessive. We further also make a labelling mistake, interpreting *america* as an organization. Figure 6 also contains such a labelling mistake, in which we fail to resolve *take part* to `participate-01`. In the example in Figure 7 the wrong sense is chosen for *fall*.

A quantitative analysis of this type of mistake shows that there is quite some room for improvement to be made by correcting these. Assuming perfect Smatch alignment of triples, 18.5% of relations are mislabelled (for instance, `ARG1` when `ARG2` would be appropriate), and 42.8% of instances are mislabelled (for example, `fall-01` when `fall-07` would be appropriate).

5 Conclusion

In this paper we wanted to investigate how feasible it is to map DRSs to AMRs. DRSs and AMRs have a lot of points in common, but there are also significant differences. We approached the problem with a three-fold strategy: lexical adaptation (changing lexical entries of the Boxer system to match AMR), a recursive translation function from DRS to AMR, and a post-processing step (needed because of the differences in verbalization and symbol labelling in AMR).

On the one hand, the overall results are perhaps disappointing. The obtained F-score does not match that of state-of-the-art semantic parsers that are trained on gold-standard AMR datasets. On the other hand, with relatively little effort reasonable output is produced. For notoriously hard constructions such as control and coordination Boxer performs well.

The question remains whether this is a promising way of producing different semantic representations (i.e., AMRs instead of DRSs). It would be interesting for future research to investigate the possibility to make Boxer's syntax-semantics interface more transparent and transform the three-step process into two phases, eliminating the need for translating DRS to AMR. Needless to say, AMR is not a replacement for DRS, as it has less expressive power, but the ability to switch between the two formats would be a welcome feature.

```
(e6 / and
 :op1 (k1 / thug
    :domain (x1 / they))
 :op2 (k2 / deserve-01
    :ARG0 x1
    :ARG1 (x2 / bullet)))
```

```
(a / and
  :op1 (t / thug
    :domain (t2 / they))
  :op2 (d / deserve-01
    :ARG0 t
    :ARG1 (b / bullet)))
```

Figure 3: They are thugs and deserve a bullet. (#111, F-score: 90.9, Boxer left, gold right)

```
(e1 / arrest-01
:ARG1 (x1 / person
  :ARG0-of (v1002 / protest-01)))
```

```
(a / arrest-01
:ARG1 (p / person :quant 1
  :ARG0-of (p2 / protest-01)))
```

Figure 4: A protester was arrested. (#710, F-score: 92.3, Boxer left, gold right)

```
(e1 / create-01
 :ARG1 (x1 / force
    :mod (s1 / country
       :name (p1002 / name
        :op1 "afghanistan")
        :wiki "afghanistan" )
    :poss (x2 / security))
 :ARG0 (x3 / and
    :op1 (x4 / organization
       :name (n3 / name
        :op1 "us")
        :wiki "United_States" )
    :op2 (x5 / coalition)))
```

```
(c3 / create-01
  :ARG0 (a / and
    :op1 (c2 / country
      :wiki "United_States"
      :name (n2 / name :op1 "US"))
    :op2 (c4 / coalition))
  :ARG1 (f / force
    :purpose (s / security)
    :mod (c / country
      :wiki "Afghanistan"
      :name (n / name
      :op1 "Afghanistan"))))
```

Figure 5: The Afghan security force was created by the US and the coalition. (#300, F-score: 90.9, Boxer left, gold right)

```
(e1 / tell-01
 :ARG0 (x1 / they)
 :ARG1 (p1 / and
    :op1 (k1 / avoid-01
     :ARG0 (x2 / she)
     :ARG1 (x3 / cafeteria))
    :op2 (k2 / take-01
     :ARG0 x2
     :ARG1 (x4 / part)
     :polarity -
     :in (x5 / homecoming)))
 :ARG2 x2)
```

```
(t / tell-01
  :ARG0 (t2 / they)
  :ARG1 (a / and
    :op1 (a2 / avoid-01
      :ARG0 s
      :ARG1 (c / cafeteria))
    :op2 (p / participate-01
      :polarity -
      :ARG0 s
      :ARG1 (h / homecoming)))
  :ARG2 (s / she))
```

Figure 6: They told her to avoid the cafeteria and not take part in homecoming. (#151, F-score: 85.0, Boxer left, gold right)

```
(e1 / fall-01
 :ARG0 (x1 / man
    :mod (s1 / innocent))
 :ARG1 (x2 / victim)
 :to (x3 / machine))
```

```
(f / fall-07
  :ARG1 (m / man
    :ARG1-of (i / innocent-01)
    :mod (a / another))
  :ARG2 (v / victimize-01
    :ARG0 (m2 / machine)
    :ARG1 m))
```

Figure 7: Another innocent man falls victim to the Machine. (#1024, F-score: 26.1, Boxer left, gold right)

Acknowledgments

We would like to thank the three anonymous reviewers for their helpful comments. We also thank Jon May for the splendid organization of the shared task. This work was partially funded by the NWO-VICI grant "Lost in Translation – Found in Meaning" (288-89-003).

References

Yoav Artzi, Kenton Lee, and Luke Zettlemoyer. 2015. Broad-coverage CCG semantic parsing with AMR. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 1699–1710.

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.

Johan Bos. 2015. Open-domain semantic parsing with boxer. In Beáta Megyesi, editor, *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 301–304.

Johan Bos. 2016. Expressive Power of Abstract Meaning Representations. *Computational Linguistics*, 42.

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.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics 2013)*, pages 121–124.

Kilian Evang, Valerio Basile, Grzegorz Chrupała, and Johan Bos. 2013. Elephant: Sequence labeling for word and sentence segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1426.

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

Hans Kamp. 1984. A Theory of Truth and Semantic Representation. In Jeroen Groenendijk, Theo M.V. Janssen, and Martin Stokhof, editors, *Truth, Interpretation and Information*, pages 1–41. FORIS, Dordrecht – Holland/Cinnaminson – U.S.A.

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

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 1375–1384.

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