Annotating semantic roles in a lexicalised grammar environment

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

Kilian Evang
CLCG
University of Groningen
The Netherlands
k.evang@rug.nl

Malvina Nissim
Linguistics and Oriental Studies
University of Bologna
Bologna, Italy
malvina.nissim@unibo.it

Abstract

Annotating text with abstract information such as semantic roles is costly. In previous efforts, such as PropBank, this process was aided with the help of syntactic trees, manually correcting automatically produced annotations. We argue that when using a lexicalised approach the annotation effort can be made simpler, avoiding the need to explicitly select two entities for each role. Our model is demonstrated by the Groningen Meaning Bank, using Combinatory Categorial Grammar as syntactic formalism, and Discourse Representation Theory as a formal semantic backbone.

1 Introduction and background

Annotating thematic roles is a time-consuming business: given an annotation scheme, for each role two entities need to be identified in the text, and the relation between them selected. This is often carried out with the help of syntactic trees and complex annotation aids. Perhaps this process can be made easier if it is considered as part of a larger semantically-oriented annotation effort. In this paper we argue that this is indeed the case.

Viewed from a simple but global perspective, annotation of thematic roles could be carried out on the surface (token) level, syntactic level, or semantic level. Perhaps, intuitively speaking, annotating semantic roles should take place at the semantic level (a logical form of some kind), because that's eventually where semantic roles belong. But reading and editing logical forms can be hard and requires extensive training for non-semanticists. Human annotation on the surface level, on the other hand, seems attractive but turns out to be a tiresome process without the aid of part-of-speech and requires sophisticated tools to select entities and specify relations between them.

There has been ample interest in semantic roles recently in the Natural Language Processing community. The main resource encoding subcategorisation frames and semantic roles over verb classes is VerbNet (Kipper Schuler, 2005). FrameNet (Baker et al., 1998) also encodes semantic roles, and it does so at a more detailed level than VerbNet, including adjuncts too, but has a much more limited coverage. NomBank (Meyers et al., 2004) provides semantic roles for nouns rather than verbs. The primary corpus annotated for semantic roles is PropBank (Palmer et al., 2005), which was annotated by hand-correcting the output of a rule-based tagger over constituency-based syntactic trees.

The evident need for joint modelling of syntactic dependencies and semantic roles has prompted a revision of PropBank for the CoNLL-2008 Shared Task on "Joint Parsing of Syntactic and Semantic Dependencies" (Surdeanu et al., 2008). One extension is the annotation of roles for the arguments of nouns as well, exploiting NomBank. The other, major, amendment is the translation of the original constituent-based structures into dependency-based ones, as a dependency grammar framework is believed to model more appropriately the syntax-semantics interface for the annotation of semantic roles (Johansson and Nugues, 2008).

Our claim is that the annotation of semantic roles is best done with the help of a *lexicalised grammati-*

cal framework. In a lexicalised grammar, verbs (and nouns) encode all their arguments inside their lexical category. This has some pleasant consequences: tokens can be easily divided into those that trigger (a finite, ordered set of) semantic roles and those that do not. Annotation then boils down to assigning the correct roles to each token. There is no need to select entities. Roles can be derived from existing resources such as VerbNet and FrameNet, depending on the desired granularity and taking into account coverage issues.

Thus, we propose a strongly lexicalised model where roles are assigned to verbs and modifiers, deriving them from external resources, and are subsequently inherited by the arguments and adjuncts directly through syntactic composition. Our experiments are implemented as part of the Groningen Meaning Bank (GMB, henceforth), a project that aims to annotate texts with formal semantic representations (Basile et al., 2012). The syntactic formalism used in the GMB is Combinatory Categorial Grammar (CCG), a lexicalised framework where syntactic categories are composed out of a few base categories (S, NP, N, PP), and slashes of complex categories indicate the direction of arguments (e.g., S\NP is a complex category looking for an noun phrase on its left to complete a sentence). The semantic formalism adopted by the GMB is Discourse Representation Theory, with a neo-Davidsonian view on event semantics.

2 Annotation Model

Semantic relations are relations between two entities, of which one is the internal and one the external entity. In the GMB semantic relations are two-place relations between discourse referents. The internal entity is usually an event, triggered by a verb; the external entity is usually triggered by a noun phrase. External entities are realised by arguments or adjuncts – annotation of roles differs with respect to whether external entities are arguments or adjuncts.

We will outline our model using the VerbNet inventory of roles for the verb *to build*. Let's first consider the annotation of roles whose external entities are introduced by arguments. In the GMB corpus various CCG categories are assigned to *build*, corresponding to different subcategorisation frames. The verb *build* is listed in two VerbNet classes: build-26.1-1 (WordNet sense 1); base-97.1 (WordNet sense 8).

Table 1 shows that *build* could be mapped to (at least) seven different VerbNet frames. However, the different CCG categories assigned to *build* already aid in disambiguating: the intransitive form S\NP maps to one VerbNet frame, the transitive form (S\NP)/NP to just three of the possible seven VerbNet frames. Whenever a CCG-category for a given verb could be mapped to more than one VerbNet frame, annotators will be presented with the relevant *roleset* (Palmer et al., 2005), i.e. the set of available role values to choose from associated to that verb usage. In the case of (S\NP)/NP, for example, Agent, Material, or Asset could be selected for the subject NP, while the object would be Product in any case.

The last column of Table 1 shows how the VN roles are inserted in the CCG categories. This, in turn, allows us to introduce the roles in the lexical DRSs for the verb. For instance, the lexical entry for the transitive form of *build* is illustrated in Figure 1. Note that VerbNet also provides the WordNet sense of a verb. This is also included in the lexical DRS as part of the symbol representing the building event (build-1). See Section 3 for the way WordNet senses can be used in the model.

build
(S\NP:Agent)/NP:Product

$$\lambda n1.\lambda n2.\lambda m.(n2@\lambda x.(n1@\lambda y.(\begin{array}{|c|}\hline e\\\hline \text{build-1(e)}\\ \text{Agent(e,x)}\\ \text{Product(e,y)}\\\hline\end{array};(m@e))))$$

Figure 1: Lexical DRS for *build*.

CCG categories corresponding to passive verb forms lack the subject NP of the corresponding active forms. Active forms are distinguished by passive forms by features on the S category. In order to map passive CCG categories to VN entries one needs to bear in mind the correspondences below:

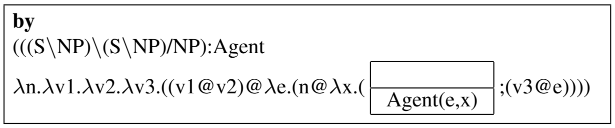
$$\mathrm{S}_{pss}\backslash\mathrm{NP}{:}\mathbf{X} \Leftrightarrow (\mathrm{S}\backslash\mathrm{NP}{:}\mathbf{Y})/\mathrm{NP}{:}\mathbf{X}$$
$$(\mathrm{S}_{pss}\backslash\mathrm{NP}{:}\mathbf{Z})/\mathrm{PP}{:}\mathbf{Y} \Leftrightarrow ((\mathrm{S}\backslash\mathrm{NP}{:}\mathbf{X})/\mathrm{PP}{:}\mathbf{Y})/\mathrm{NP}{:}\mathbf{Z}$$

This is how roles are assigned to arguments in the annotation model. For the roles that are introduced

Table 1: Mapping VerbNet roles to CCG categories, for *build*.

Category	**Class**	**Sense**	**VerbNet frame**	**Enhanced CCG category**
S\NP	build-26.1	1	Agent V	S\NP:agent
(S\NP)/NP	build-26.1	1	Agent V Product	(S\NP:agent)/NP:product
	build-26.1	1	Material V Product	(S\NP:material)/NP:product
	build-26.1-1	1	Asset V Product	(S\NP:asset)/NP:product
((S\NP)/PP)/NP	build-26.1	1	Agent V Product {from} Material	((S\NP:agent)/PP:material)/NP:product
	build-26.1-1	1	Agent V Product {for} Asset	((S\NP:agent)/PP:asset)/NP:product
	base-97.1	8	Agent V Theme {on} Source	((S\NP:agent)/PP:source)/NP:theme

by adjuncts we need a different strategy. In CCG, adjuncts are represented by categories of the form X/X or X\X, where X is any CCG category, possibly enhanced with further subcategorisation information (for instance in the case of prepositions). This will allow us to assign roles at the token level. This idea is shown in Figure 2 for a preposition (VP modifier).



by

(((S\NP)\(S\NP)/NP):Agent

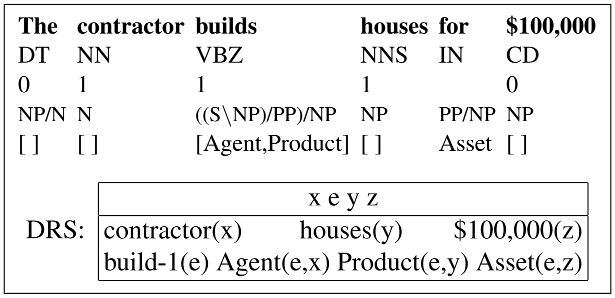
λn.λv1.λv2.λv3.((v1@v2)@λe.(n@λx.(Agent(e,x) ;(v3@e))))

Figure 2: Lexical DRS for *by*.

It is important to see that, in this annotation model, semantic roles are annotated at the token level. Given a set of tokens corresponding to a sentence, each token is associated with an ordered, possibly empty, set of tokens. The number of elements in this set is determined by the CCG category. Categories corresponding to adjuncts introduce one role, the number of roles for categories associated with verbs is determined by the number of arguments encoded in the CCG category. This makes annotation not only easier, it also makes it more flexible, because one could even annotate correct roles for a clause whose syntactic analysis is incorrect.

3 Implementation

The GMB implements a layered approach to annotation. On the token level, there are separate layers, each with its own tag-set, for part-of-speech, named entities, numeral expressions, lexical categories, word senses, among others (Figure 3). These layers all contribute to the construction of the semantic representation of the sentence, and eventually that of a text, in the form a DRS. For semantic roles of VerbNet a further annotation layer is



The	**contractor**	**builds**	**houses**	**for**	**$100,000**
DT	NN	VBZ	NNS	IN	CD
0	1	1	1		0
NP/N	N	((S\NP)/PP)/NP	NP	PP/NP	NP
[ ]	[ ]	[Agent,Product]	[ ]	Asset	[ ]

DRS:

x e y z

contractor(x) houses(y) $100,000(z)

build-1(e) Agent(e,x) Product(e,y) Asset(e,z)

Figure 3: Annotation layers in the GMB and corresponding semantic representation.

added. Note that for different inventory of roles, such as FrameNet, a further annotation layer could be included (Bos and Nissim, 2008). As we have shown in the previous section, the roles turn up in the DRS for the sentence, following the compositional semantics determined by the syntactic analysis, as two-place relation between two discourse referents (see Figure 3).

The manual annotation could be performed in three possible modes. The *open* mode lets the annotator choose from all possible VerbNet frames available for a given verb. In a *restricted* mode, the annotator can choose to activate specific constraints which limit the number of frames to choose from. For example, by activating the constraint relative to the syntactic category of the verb, for instance (S\NP)/NP, the annotator could reduce the number of possible frames for *to build* from seven to just three (see Table 1). Another constraint could be the WordNet sense: in the GMB, verb sense disambiguation is dealt with by a separate layer using the senses of WordNet, and WordNet senses are also used in VerbNet. Using the WordNet constraint, only VerbNet frames associated to a given WordNet sense would be available to choose from. For example, if sense 8 of *to build* is selected there is only one option available (see Table 1). Alternatively, the WordNet sense could be used for detecting a possible error — for example if "source" is used in combination with sense 1 of *to build*, a warning should be issued as "source" can only be used with sense 8. In the *automatic* mode, the system will produce the annotation automatically on the basis of the correspondences and constraints which we have described, and the human annotator will be able to subsequently amend it through the GMB annotation interface. Whenever there is more than one option, such as assigning the appropriate VerbNet frame to an instance of build with category (S\NP)/NP, choice strategies must be devised (see Section 4).

4 Further Issues

There are a couple of further issues that need to be addressed. First, the choice of roleset depends on the sense assigned to a verb (or noun). In the GMB, word senses and roles are implemented by two different annotation layers. The question remains whether to permit inconsistencies (supported by a system of warnings that notices the annotator might such contradictions arise) or instead implement a system that constrains the choice of roleset on the basis of the selected word sense.

As we have seen, and as it is also noted by (Palmer et al., 2005), the same verb can be listed more than once with the same subcategorisation frame to which are however associated different roles. While in *open* and *restricted* modes the annotator will select the appropriate one, in *automatic* mode decision strategies must be devised. Another issue is to do with missing frames in VerbNet, such as for *build-8* with a NP V PP structure as in "He also seeks to build on improvements". An appropriate frame, such as Agent V Theme or Agent V Source, does not exist in VerbNet for *to build*, unlike e.g. for *to rely*. To address such cases, the interface should also let annotators choose from the whole inventory of VerbNet frames.

In the CoNLL 2008 shared task, data from NomBank is integrated with PropBank to get a wider range of arguments to be annotated for semantic roles, including thus nouns beside verbs. The lexicalised framework we have presented here can easily be extended to cover NomBank data as well.

Finally, this annotation model also has consequences for predicting semantic roles by machines. This is because, in a lexicalised framework such as the one that we propose, the process of semantic role labelling is essentially transformed to a classification task on tokens. Whether this could lead to better performance in semantic role labelling is a question left for future research.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, pages 86–90. ACL.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3196–3200, Istanbul.

Johan Bos and Malvina Nissim. 2008. Combining Discourse Representation Theory with FrameNet. In R. Rossini Favretti, editor, *Frames, Corpora, and Knowledge Representation*, pages 169–183. Bononia University Press, Bologna.

Richard Johansson and Pierre Nugues. 2008. Dependency-based semantic role labeling of propbank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 69–78, Stroudsburg, PA, USA. ACL.

Karin Kipper Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The nombank project: An interim report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA. ACL.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, March.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177, Manchester, England, August.