Towards Generating Text from Discourse Representation Structures

Valerio Basile
Humanities Computing
University of Groningen
v.basile@rug.nl

Johan Bos
Humanities Computing
University of Groningen
johan.bos@rug.nl

Abstract

We argue that Discourse Representation Structures form a suitable level of language-neutral meaning representation for micro planning and surface realisation. DRSs can be viewed as the output of macro planning, and form the rough plan and structure for generating a text. We present the first ideas of building a large DRS corpus that enables the development of broad-coverage, robust text generators. A DRS-based generator imposes various challenges on micro-planning and surface realisation, including generating referring expressions, lexicalisation and aggregation.

1 Introduction

Natural Language Generation, NLG, is often viewed as a complex process comprising four main tasks (Bateman and Zock, 2003): i) **macro planning**, building an overall text plan; ii) **micro planning**, selecting referring expressions and appropriate content words; iii) **surface realisation**, selection of grammatical constructions and linear order; and iv) **physical presentation**, producing final articulation and layout operations. Arguably, the output of the macro planning component in an NLG system is some sort of abstract, language-neutral representation that encodes the information and messages that need to be conveyed, structured by rhetorical relations, and supported by information that is presupposed to be common ground.

We argue that the Discourse Representation Structures (DRSs) from Discourse Representation Theory (Kamp, 1984) form an appropriate representation for this task. This choice is driven by both theoretical and practical considerations:

- DRT, being a theory of analysing meaning, is by principle language-neutral;
- Many linguistic phenomena are studied in the framework provided by DRT;
- DRT has a model-theoretical backbone, allowing applications to perform logical inferences with the aid of theorem provers.

As a matter of fact, DRT has means to encode presupposed information in a principled way (Van der Sandt, 1992), and connections with rhetorical relations are spelled out in detail (Asher, 1993). Moreover, the formal integration of DRS with named entities, thematic roles and word senses is natural.

These are, mostly, purely theoretical considerations. But in order to make DRSs a practical platform for developing NLG systems a large corpus of text annotated with DRSs is required. Doing this manually is way too costly. But given the developments in (mostly statistical) parsing of the last two decades we are now in a position to use state-of-the-art tools to semi-automatically produce gold (or nearly gold) standard DRS-annotated corpora.

Such a resource could form a good basis to develop (statistical) NLG systems, and this thought is supported by current trends in broad-coverage NLG components (Elhadad and Robin, 1996; White et al., 2007), that take deep semantic representations as starting points for surface realisation. The importance of a multi-level resource for generation is underlined by Bohnet et al. (2010), who feel the lack of such a resource is hampering progress in the field.

In this paper we show how we are building such a corpus (SemBank, Section 2), what the exact nature of the DRSs in this corpus is, and what phenomena are covered (Section 3). We also illustrate what challenges it poses upon micro planning and surface realisation (Section 4). Finally, in Section 5, we discuss how generating from DRSs relates to the traditional NLG pipeline.

2 The Groningen SemBank

Various semantically annotated corpora of reasonable size exist nowadays: PropBank (Palmer et al.,

Proceedings of the 13th European Workshop on Natural Language Generation (ENLG), pages 145–150,
Nancy, France, September 2011.

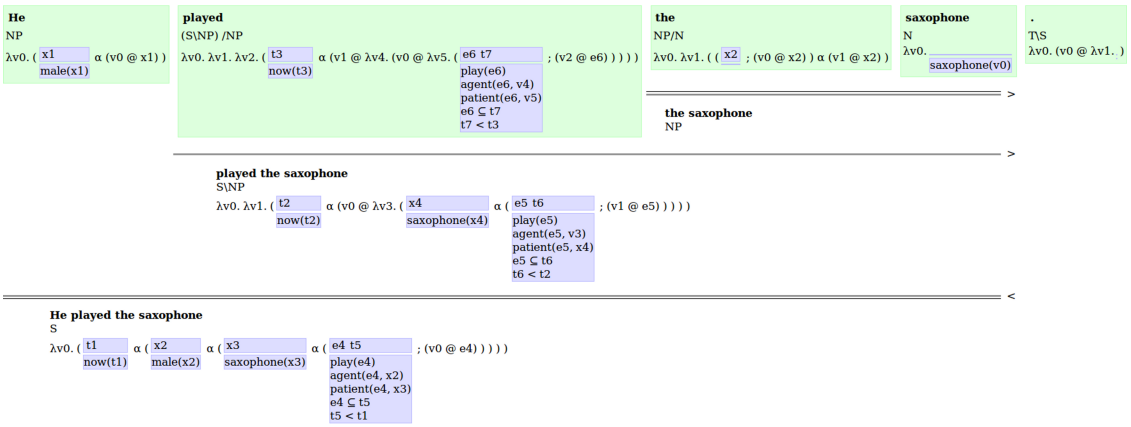

Figure 1: Screenshot of SemBank's visualisation tool for the syntax-semantics interface combining CCG and DRT.

2005), FrameNet (Baker et al., 1998), the Penn Discourse TreeBank (Prasad et al., 2005), and several resources developed for shared tasks such as CoNNL and SemEval. Annotated corpora that combine various levels of annotation into one formalism hardly exist. A notable exception is OntoNotes (Hovy et al., 2006), combining syntax (Penn Treebank style), predicate argument structure (based on PropBank), word senses, and coreference. Yet all of these resources lack a level comprising a formally grounded "deep" semantic representation that combines various layers of linguistic annotation.

Filling this gap is exactly the purpose of SemBank. It provides a collection of semantically annotated texts with **deep rather than shallow semantics**. Its goal is to **integrate phenomena instead of covering single phenomena** into one formalism, and representing **texts, not sentences**. SemBank is driven by linguistic theory, using CCG, Combinatory Categorial Grammar (Steedman, 2001), for providing syntactic structure, employing (Segmented) Discourse Representation Theory (Kamp, 1984; Asher and Lascarides, 2003) as semantic framework, and first-order logic as a language for automated inference tasks.

In our view, a corpus developed primarily for research purposes must be widely available to researchers in the field. Therefore, SemBank will only consists of texts which distribution isn't subject to copyright restrictions. Currently, we focus on English newswire text from an American newspaper whose articles are in the public domain. In the future we aim to cover other text genres, possibly integrating resources from the Open American National Corpus (Ide et al., 2010). The plan is to release a stable version of SemBank in regular intervals, and to provide open access to the development version.

The linguistic levels of SemBank are, in order of analysis depth: part of speech tags (Penn tagset); named entities (roughly based on the ACE ontology); word senses (WordNet); thematic roles (VerbNet); syntactic structure (CCG); semantic representations, including events and tense (DRT); rhetorical relations (SDRT). Even though we talk about different levels here, they are all connected to each other. We will show how in the following section.

Size and quality are factors that influence the usefulness of annotated resources. As one of the things we have in mind is the use of statistical techniques in NLG, the corpus should be sufficiently large. However, annotating a reasonably large corpus with gold-standard semantic representations is obviously a hard and time-consuming task. We aim to provide a trade-off between quality and quantity, with a process that improves the annotation accuracy in each periodical stable release of SemBank.

This brings us to the method we employ to construct SemBank. We are using state-of-the-art tools for syntactic and semantic processing to provide a rough, first proposal of semantic representation for a text. Among other tools, the most important are the C&C parser (Clark and Curran, 2004) for syntactic analysis, and Boxer (Bos, 2008) for semantic analysis. This software, trained and developed on the Penn Treebank, shows high coverage for texts in the newswire domain (up to 98%), is robust and fast, and therefore suitable for this task.

The output of these tools are corrected by crowd-

Table 1: Illustration of linguistic information integration in SemBank

Level	Theory/Source	Internal DRS Encoding
semantics	DRT (Kamp and Reyle, 1993)	`drs(...,...)`
named entity	ACE	`named(X,'Clinton',per)`
thematic roles	VerbNet (Kipper et al., 2008)	`rel(E,X,'Agent')`
word senses	WordNet (Fellbaum, 1998)	`pred(X,loon,n,2)`
rhetorical relations	SDRT (Asher and Lascarides, 2003)	`rel(K1,K2,elaboration)`

sourcing methods, comprising (i) a group of experts that are able to propose corrections at various levels of annotation in a wiki-based fashion; and (ii) a group of non-experts that provide information for the lower levels of annotation decisions by way of a *Game with a Purpose*, similar to the successful Phrase Detectives (Chamberlain et al., 2008) and Jeux de Mots (Artignan et al., 2009).

3 Discourse Representation Structures

A DRS comprises two parts: a set of discourse referents (the entities introduced in the text), and a set of conditions, describing the properties of the referents and the relations between them. We adopt well-known extensions to the standard theory to include rhetorical relations (Asher, 1993) and presuppositions (Van der Sandt, 1992). DRSs are traditionally visualised as boxes, with the referents placed in the top part, and the DRS conditions in the bottom part. The convention in SemBank is to sort the discourse referents into entities (variables starting with an x), events (e), propositions (p), temporalities (t), and discourse segments (k), as Figure 2 shows.

The DRS conditions can be divided into basic and complex conditions. The basic conditions are used to describe names of discourse referents (`named`), concepts of entities (`pred`), relations between discourse referents (`rel`), cardinality of discourse referents denoting sets of objects (`card`), or to express identity between discourse referents (=). The complex conditions introduce embedded DRSs: implication ($\Rightarrow$), negation ($\neg$), disjunction ($\vee$), and modalities ($\Box$, $\Diamond$). DRSs are thus of recursive nature, and the embedding of DRSs restrict the resolution of pronouns (and other anaphoric expressions), which is one of the trade mark properties of DRT.

The aim of SemBank is to provide fully resolved semantic representations. Obviously, natural language expressions can be ambiguous and picking the most likely interpretation isn't always straightforward: Some pronouns have no clear antecedents, word senses are often hard to distinguish, and scope orderings are sometimes vague. In future work this might give rise to adding some underspecification mechanisms into the formalism.

DRSs are formal structures and come with a model-theoretic interpretation. This interpretation can be given directly (Kamp and Reyle, 1993) or via a translation into first-order logic (Muskens, 1996). This is interesting from a practical perspective, because it permits the use of efficient existing inference engines developed by the automated deduction community. Applying logical inference can play a role in tasks surrounding NLG (e.g., summarisation, question answering, or textual entailment), but also dedicated components of NLG systems, such as generating definite descriptions, which requires checking contextual restrictions (Gardent et al., 2004).

Figure 1 illustrates how SemBank provides the compositional semantics of each sentence in the text in the form of a CCG derivation. Here each token is associated with a supertag (a lexical CCG category) and its corresponding lexical semantics, a partial DRS. The CCG derivation, a tree structure, shows the compositional semantics in each step of the derivation, with the aid of the λ-calculus (the @ operator denotes function application).

Table 1 shows how the various levels of annotation are integrated in DRSs. Thematic roles (VerbNet) are implied by the neo-Davidsonian event semantics employed in SemBank, and are represented as two-place relations. The named entity types form part of the basic DRS condition for names, and Word senses (WordNet) are represented as a feature on the one-place conditions for nouns, verbs and modifiers. Rhetorical relations are already part and parcel of SDRT. Hence, SemBank provides all these different layers of information within a DRS. Figure 2 shows an SDRS for a small text of SemBank.

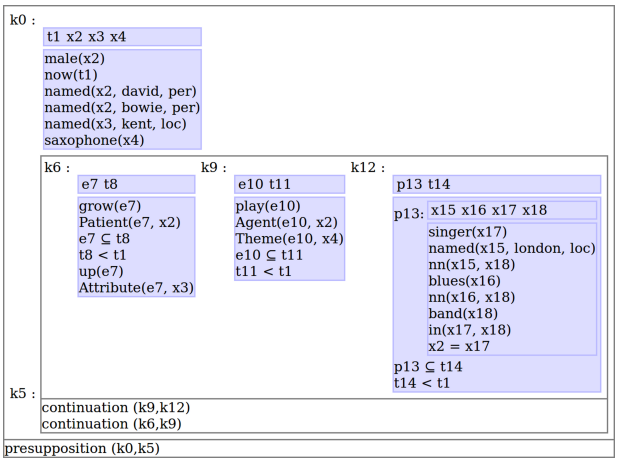

Figure 2: SDRS for the text "David Bowie grew up in Kent. He played the saxophone. He was a singer in London blues bands", as shown in SemBank.

4 Challenges for Text Generation

We believe that taking DRS as the basis for NLG will introduce not only variants of known problems, but will impose many new challenges. Here we focus on just three of them: generating referring expressions, lexicalisation, and aggregation.

4.1 Generating referring expressions

Viewed from a formal perspective, DRT is said to be a *dynamic* theory of semantics: the interpretation of an embedded DRS depends on the interpretation of the DRSs that subordinate it — either be sentence-internal structure, or by the structure governed by rhetorical relations. A case in point is the treatment of anaphoric expressions including pronouns, proper names, possessive constructions and definite descriptions.

In DRT, anaphoric expressions are resolved to a suitable antecedent discourse referent. Proper names and definite descriptions are too, but if finding a suitable antecedent fails then a process usually referred to as *presuppositional accommodation* introduces the semantic material of the anaphoric expression on an accessible level of DRS (Van der Sandt, 1992). The result of this process yields a DRS in which all presupposed information is explicitly distinguished from asserted information. This gives rise to an interesting challenge for NLG.

A DRS corresponding to a discourse unit will contain free variables for semantic material that is presupposed or has been linked to the preceeding context. On encountering such a free variable denoting an entity, the generator has a couple of choices in the way it can lexicalise it: as a quantifier, pronoun, proper name, or definite description. Even though the DRS context may provide information on names and properties assigned to this free variable, we expect it will be non-trivial to decide what properties to include in the corresponding expression. Text coherence probably plays an important role here, but whether thematic roles and rhetorical relations will be sufficient to predict an appropriate surface form remains a subject for future research. It is also interesting to explore the insights from approaches dedicated to generating referring expressions using logical methods (van Deemter, 2006; Gardent et al., 2004) with robust surface realisation systems.

4.2 Aggregation

Coordinated noun phrases are known to be potentially ambiguous between distributive and collective interpretations. A simple DRT analysis for the distributive interpretation yields two possible ways to generate strings: one where the noun phrases are coordinated within one sentence, and one where the noun phrases involved are generated in separate sentences. For instance, the DRSs corresponding to "Deep Purple and Pink Floyd played at a charity show" (with a distributive interpretation) and "Deep Purple played at a charity show, and Pink Floyd played at a charity show", would be equivalent. This is due to copying semantic material in the compositional process of computing the meaning of the coordinated noun phrase "Deep Purple and Pink Floyd". (Note that the *collective* reading, as in "Deep Purple and Pink Floyd played together at a charity show" would not involve copying semantic material, and would result in a different DRS, with a different interpretation.) It is the task of the aggregation process to pick one of these realisations, as discussed by White (2006). Doing this from the level of DRS poses an interesting challenge, because one would need to recognise that such an aggregation choice is possible in the first place. Alternatively, instead of copying, one could use an explicit operator that signals a distributive reading of a plural noun phrase, for instance as suggested by Kamp and Reyle (1993). Arguably, this is required anyway to adequately represent sentences such as "Both Deep Purple and Pink Floyd played at a charity show".

4.3 Lexicalisation

The predicates (the one-place relations) found in a DRS correspond to concepts of an hierarchical ontology. Time expressions and numerals have canonical representations in SemBank. The representation for noun and verb concepts are based on the synonym sets provided by WordNet (Fellbaum, 1998). A WordNet synset can be referred to by its internal identifier, or by any of its member word-sense pairs. For instance, synset `102999757` is composed of the noun–sense pairs `strand-3`, `string-10`, and `chain-10`. The lexicalisation challenge is to select the most suitable word out of these possibilities. Local context might help to choose: a "string of beads" is perhaps better than a "chain of beads" or "strand of beads". As another example, consider the synset {`loon-2`,`diver-3`} representing the concept for a kind of bird. American birdwatchers would use the noun "loon", whereas in Britain "diver" would be preferred to name this bird.

5 Discussion

In the DRT-based framework that we propose for generating text, the issue arises *where* in the traditional NLG pipeline DRSs play a role. In the introduction of this paper we suggested that DRSs would be output by the macro planner, and hence fed as input to the micro planner. On the one hand this makes sense, as in a segmented DRS all content to be generated is present and the rhetorical structure is given explicitly. But then the question remains whether the theoretical distinction between micro planning and surface realisation really works in practice or would just be counter-productive. Perhaps a revised architecture tailored to DRS generation should be tried instead. This issue is closely connected to the level of semantic granularity that one would like to see in a DRS. We illustrate this by four examples:

- **pronouns** — we have made a particular proposal using free variables, but we could also have followed Kamp and Reyle (1993), introducing explicit referents for pronouns;
- **distributive noun phrases** — as the discussion in Section 4.2 shows, it is unclear which represention for distributive noun phrases would be most suitable for the purpose of sentence planning;
- **sentential and verbal complements** — should there be a difference in meaning representation between "Tim expects to win" and "Tim expects that he will win"?
- **active vs. passive voice** — should a meaning representation reflect the difference between active and passive sentences?

At this moment, it is not clear whether one wants a more abstract DRS and give more freedom to sentence planning, or a more specific DRS restricting the number of sentential paraphrases of its content. Perhaps even an architecture permitting both extremes would be feasible, where the task of micro planning would be to add more constraints to the DRS until it is specific enough for the surface realisation component to generate text from it. It is even thinkable that such a planning component would take over some tasks of the macro planner, making the distinction between the two fuzzier.

A final point that we want to raise is a possible role that inference can play in our framework. DRSs can be structurally different, yet logically equivalent. This could influence the design of a generation system and have a positive impact on its output. For instance, it would be thinkable to equip the NLG system with a set of meaning-preserving tranformation rules that change the structure of a DRS, consequently producing different surface forms.

6 Conclusion

SemBank provides an annotated corpus combining shallow with formal semantic representations for texts. The development version is currently available online with more than 60,000 automatically annotated texts; the release of a first stable version comprising ca. 1,000 texts is planned later this year. We expect SemBank to be a useful resource to make progress in robust NLG. Using DRSs as a basis for generation poses new challenges, but also could offer fresh perspectives on existing problems in NLG.

Acknowledgments

We are grateful to Michael White, who provided us with useful feedback to the idea of using a DRS corpus for developing and training text generation systems. We also would like to thank the three anonymous reviewers of this article. They gave extremely valuable comments that considerably improved our paper.

References

Guillaume Artignan, Mountaz Hascoët, and Mathieu Lafourcade. 2009. Multiscale visual analysis of lexical networks. *Information Visualisation, International Conference on*, 0:685–690.

N. Asher and A. Lascarides. 2003. *Logics of conversation*. Studies in natural language processing. Cambridge University Press.

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

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Proceedings of the Conference*, pages 86–90, Université de Montréal, Montreal, Quebec, Canada.

John Bateman and Michael Zock. 2003. Natural Language Generation. In R. Mitkov, editor, *Oxford Handbook of Computational Linguistics*, chapter 15, pages 284–304. Oxford University Press, Oxford.

Bernd Bohnet, Leo Wanner, Simon Mille, and Alicia Burga. 2010. Broad coverage multilingual deep sentence generation with a stochastic multi-level realizer. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 98–106.

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.

John Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2008. Addressing the Resource Bottleneck to Create Large-Scale Annotated Texts. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 375–380. 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.

Michael Elhadad and Jacques Robin. 1996. An overview of SURGE: a reusable comprehensive syntactic realization component. Technical report.

Christiane Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

Claire Gardent, Hélène Manuélian, Kristina Striegnitz, and Marilisa Amoia. 2004. Generating definite descriptions: Non-incrementality, inference and data. In Thomas Pechmann and Christopher Habel, editors, *Multidisciplinary approaches to language production*, pages 53–85. Walter de Gruyter, Berlin.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, Stroudsburg, PA, USA.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: a community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73, Stroudsburg, PA, USA.

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.

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.

Reinhard Muskens. 1996. Combining Montague Semantics and Discourse Representation. *Linguistics and Philosophy*, 19:143–186.

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

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The Penn Discourse TreeBank as a resource for natural language generation. In *Proc. of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32.

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

Kees van Deemter. 2006. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

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

Michael White, Rajakrishnan Rajkumar, and Scott Martin. 2007. Towards broad coverage surface realization with CCG. In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*.

Michael White. 2006. Efficient realization of coordinate structures in combinatory categorial grammar. *Research on Language & Computation*, 4:39–75.