ISBN 978-83-60810-27-9
ISSN 1896-7094

Generation of First-Order Expressions from a Broad Coverage HPSG Grammar

Ravi Coote and Andreas Wotzlaw
Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE
Neuenahrer Str. 20, 53343 Wachtberg, Germany
{ravi.coote, andreas.wotzlaw}@fkie.fraunhofer.de

***Abstract*—This paper describes an application for computing first-order semantic representations of English texts. It is based on a combination of hybrid shallow-deep components arranged within the middleware framework Heart of Gold. The shallow-deep semantic analysis employs Robust Minimal Recursion Semantics (RMRS) as a common semantic underspecification formalism for natural language processing components. In order to compute efficiently first-order representations of the input text, the intermediate RMRS results of the shallow-deep analysis are transformed into the dominance constraints formalism and resolved by the underspecification resolver UTool. First-order expressions can serve as a formal knowledge representation of natural text and thus can be utilized in knowledge engineering or textual reasoning. At the end of this paper, we describe their application for recognizing textual entailment.**

***Index Terms*—recognizing textual entailment; logical inference; HPSG-based text analysis; first-order logics**

I. Introduction

MANY applications depend on a formal representation of natural language sentences in form of *first-order logic* (FOL). For instance, in recognizing textual entailment [1] the approaches based on logical inference depend on FOL formulas as a semantic representation of the input texts. Furthermore, enhanced first-order knowledge representations of natural language sentences can be accomplished by integrating knowledge from resources like the lexical base *WordNet* [2] or the ontological database *YAGO* [3] into FOL formulas. Other applications can be found in the area of *knowledge engineering* and *information integration* (see, e.g., [4]).

Basically, for a production of fine-grained FOL expressions a broad coverage of syntactic structures and English words is preferable. To this end, the *English Resource Grammar* (ERG, see [5]), a broad-coverage, linguistically precise *Head-driven Phrase Structure Grammar* (HPSG) of English can be used. ERG utilizes *Minimal Recursion Semantics* (MRS, see [6]) as a formalism for building scope underspecified semantic representations from HPSG grammars.

Combining deep parsing techniques with shallow ones can make parsing processes more robust, i.e., less error-prone and faster [7]. For that reason, we use *Heart of Gold* (HOG, see [7]), a framework for combining NLP components like, e.g., shallow statistical parsers, named entity recognizers, and deep syntactic parsers.

Neither literature nor implemented systems producing FOL from an HPSG grammar with RMRS as semantic formalism could be found so far. For that reason, we decided to build such a system which we present in this paper.

Related work: Other approaches of wide coverage syntactic and semantic analysis with FOL output as input for textual entailment were presented in [8]. However, in contrast to our approach they are not based on a purely linguistically motivated grammar formalism with a broad coverage like, e.g., ERG. Furthermore, our formula generation system has successfully been employed as underlying analysis basis in a framework for recognizing textual entailment (see [9]).

II. Semantic Representation Formalisms

In this section we describe shortly the semantic formalisms MRS and RMRS on which our application builds.

Minimal Recursion Semantics: Scope underspecification is a well-known technique in computational semantics of natural language [10]. MRS is a description language over formulas of FOL languages with *generalized quantifiers*. For instance, the sentence *"Every wizard acts in a circus"* illustrates the well-known problem of scopal ambiguity. Is it one and the same circus in which every wizard acts or are there possibly several different circuses in which the wizards act? Thus, the sentence has two scopal *readings* which can be represented by the following FOL formulas:

$$a(x_9, circus(x_9), every(x_5, wizard(x_5), and(act(e_2, x_5), in(e_2, x_9)))) \quad (1)$$

$$every(x_5, wizard(x_5), a(x_9, circus(x_9), and(act(e_2, x_5), in(e_2, x_9)))). \quad (2)$$

MRS allow multiple formulas, which differ only in their scopal configuration like, e.g., (1) and (2), to be expressed with exactly one single compact formula. To achieve this, in MRS predicates of the formulas are decoupled from each other by removing any nesting of predicates, assigning different *labels* $l_1, l_2, ...$ to them, and adding *holes* $h_1, h_2, ...$ to scope relevant predicates. A single scope underspecified representation of (1) and (2) above can then be given as an MRS by

$$\begin{aligned} < \{&l_3 : every(x_5, h_6, h_4), l_7 : wizard(x_5), \\ &l_8 : act(e_2, x_5), l_8 : in(e_{10}, e_2, x_9), \\ &l_{11} : a(x_9, h_{13}, h_{12}), l_{14} : circus(x_9)\}, \\ \{&h_6\ qeq\ l_7,\ \ h_{13}\ qeq\ l_{14}\} > \end{aligned} \quad (3)$$

More specifically, (3) contains a set of labeled (decoupled) first-order predicates ($l_3 : every, l_7 : wizard, ...$) with holes ($h_6, h_4, ...$) and a set of *qeq-constraints* ($h_6\, qeq\, l_7, ...$). A *qeq-constraint* states a directive enforcing a particular label l to be in the scope of a particular hole h. In (3), the first qeq-constraint enforce label l_7 to be in the scope of h_6. Scope specified formulas are obtained by assigning, or *plugging*, labeled predicates to holes in a manner that is consistent with the qeq-constraints. For instance, (1) is obtained by plugging l_7 into h_6, l_{14} into h_{13}, l_3 into h_{12}, and l_8 into h_4. Such formalized assignments are called also pluggings.

Robust Minimal Recursion Semantics: RMRS is a generalization of MRS. It can not only be underspecified for scope as MRS, but also partially specified, e.g., when some parts of the text cannot be resolved by a given NLP component. Furthermore, in RMRS due to possible lack of morphological analysis, predicates are allowed to lack for their arguments. Hence, it can be used as a semantic representation formalism of shallow NLP components. HOG supports integration of shallow NLP components by using RMRS as an exchange format. Additionally, RMRS defines an in-group relation ing which describes conjunctively connected labels, e.g., the in-group relation $\{h8\ ing\ h10001\}$ in Figure 2. For a detailed description of MRS and RMRS see [6] and [11], respectively.

III. Generation of FOL Formulas

The application for generation of FOL formulas illustrated in Figure 1 consists of two parts:

a) Semantic analysis with the shallow-deep approach realized by HOG, and
b) Resolving of RMRS realized by UTool.

Aditionally, a graphical user interface (GUI) enclosing these parts was developed to control the analysis process and to inspect the results of the analysis components.

HOG is configured to control the overall workflow of the hybrid shallow and deep analysis. In particular, it transforms and transports data among various NLP components of the application. First, HOG processes the input text and computes its semantic representation as RMRS. Afterwards, the generation of *pluggings* representing readings of the RMRS is performed with the underspecification solver *UTool* [12]. Finally, the FOL formulas are constructed out of the computed pluggings and the original RMRS. In the following the processing steps are described in more detail.

Syntactic-semantic analysis: Shallow parsing techniques are used to retrieve superficial information from the sentences without a deep structure. In our system shallow parsing begins with tokenization by JTok (distributed with HOG) which splits the sentence by its tokens. Those tokens are passed to the statistical part-of-speech tagger TnT [13]. In parallel to these two components, SProUT [14], a finite state machine named entity recognizer, prepares information about named entities found. The results from TnT and SProUT are merged together via XML-transformation from HOG and prepared for input to the deep HPSG parser PET [15]. In that manner, PET is supplied with information for words which possibly are not contained in the deep HPSG lexicon (e.g., unusual named entities). Afterwards, PET parses the pre-annotated input text by employing the HPSG grammar ERG. With the help of this processing step, a fully syntactic annoted phrase structure is computed, from which predicate argument structures can directly be established. During the deep HPSG analysis, PET composes (robust minimal recursion) semantics according to the semantic algebra for HPSG grammars [16]. According to the analysis procedure described above, HOG produces the RMRS in Figure 2 for the following example sentence:

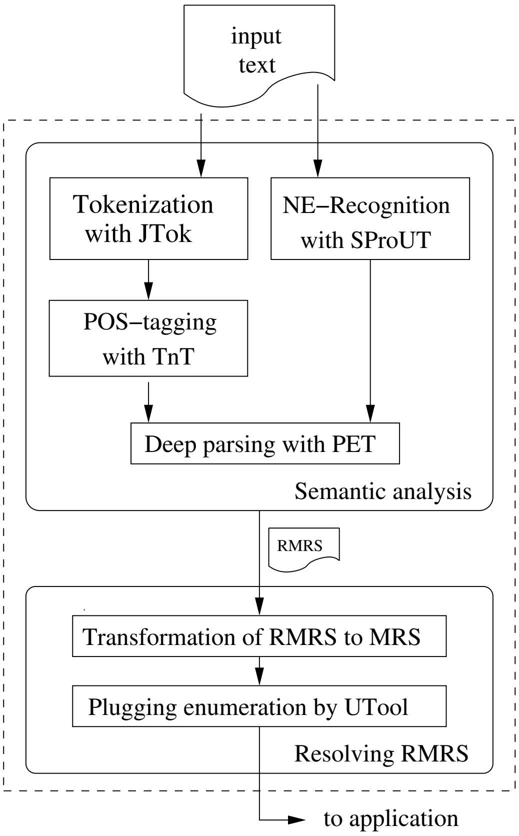

Fig. 1. Overall system for formula generation.

```
A wizard acts in a circus show in Paris.
```

Resolving RMRS: Unfortunately, a sentence with n quantifiers can have up to $n!$ readings [6], e.g., RMRS in Figure 2 with the quantifiers a_q, $udef\_q$, and $proper\_q$ has 4! scopal readings. More adverse is that about 8% of the sentences of the Rondane treebank provided with ERG have more than 100,000 readings according to the ERG analyses, and about 4% have more than one million readings [17]. Thus, it is required to enumerate all readings efficiently and to eliminate those which are logically equivalent. These tasks are performed by UTool in time of $O(n^2)$ per solved form (see [18]).

As in Section II described, for MRS the scopal readings are obtained by plugging labels l to holes h. RMRS has to be resolved analogously. However, in its standard configuration, UTool resolves only MRS. Therefore, by following the in-

$$\begin{bmatrix}
\text{TEXT} & \\
\text{TOP} & h1 \\
\text{RELS} & \left\{ \begin{array}{l}
\begin{bmatrix} \textit{\_a\_q} & \\ \text{LBL} & h3 \\ \text{ARG0} & x5 \\ \text{RSTR} & h6 \\ \text{BODY} & h4 \end{bmatrix}
\begin{bmatrix} \textit{\_wizard\_n} & \\ \text{LBL} & h7 \\ \text{ARG0} & x5 \end{bmatrix}
\begin{bmatrix} \textit{\_act\_v} & \\ \text{LBL} & h8 \\ \text{ARG0} & e2 \\ \text{ARG1} & x5 \end{bmatrix} \\
\begin{bmatrix} \textit{\_in\_p} & \\ \text{LBL} & h10001 \\ \text{ARG0} & e10 \\ \text{ARG1} & e2 \\ \text{ARG2} & x9 \end{bmatrix}
\begin{bmatrix} \textit{\_a\_q} & \\ \text{LBL} & h11 \\ \text{ARG0} & x9 \\ \text{RSTR} & h13 \\ \text{BODY} & h12 \end{bmatrix}
\begin{bmatrix} \textit{compound\_rel} & \\ \text{LBL} & h14 \\ \text{ARG0} & e16 \\ \text{ARG1} & x9 \\ \text{ARG2} & x15 \end{bmatrix} \\
\begin{bmatrix} \textit{udef\_q\_rel} & \\ \text{LBL} & h17 \\ \text{ARG0} & x15 \\ \text{RSTR} & h19 \\ \text{BODY} & h18 \end{bmatrix}
\begin{bmatrix} \textit{\_circus\_n} & \\ \text{LBL} & h20 \\ \text{ARG0} & x15 \end{bmatrix}
\begin{bmatrix} \textit{\_show\_n} & \\ \text{LBL} & h10002 \\ \text{ARG0} & x9 \\ \text{ARG1} & u21 \end{bmatrix} \\
\begin{bmatrix} \textit{\_in\_p} & \\ \text{LBL} & h10003 \\ \text{ARG0} & e23 \\ \text{ARG1} & e2 \\ \text{ARG2} & x22 \end{bmatrix}
\begin{bmatrix} \textit{proper\_q\_rel} & \\ \text{LBL} & h24 \\ \text{ARG0} & x22 \\ \text{RSTR} & h26 \\ \text{BODY} & h25 \end{bmatrix}
\begin{bmatrix} \textit{named\_rel} & \\ \text{LBL} & h27 \\ \text{ARG0} & x22 \\ \text{CARG} & \text{Paris} \end{bmatrix}
\end{array} \right\} \\
\text{HCONS} & \{h6 \text{ qeq } h7, h13 \text{ qeq } h14, h19 \text{ qeq } h20, h26 \text{ qeq } h27\} \\
\text{ING} & \{h8 \text{ ing } h10001, h8 \text{ ing } h10003, h14 \text{ ing } h10002\}
\end{bmatrix}$$

Fig. 2. RMRS of shallow deep analysis of the example sentence as attribute value matrix.

structions from [11], we developed an XSLT-based procedure for transforming RMRS into MRS to make the output RMRS of HOG suitable for resolving with UTool. Afterwards, UTool can automatically compute the assignments according to which the labels should be linked to holes. In particular, the following procedure is applied (see [17] for details):

1) Translation of MRS into *dominance constraints* [17], a closely related underspecification formalism, and
2) Efficient enumeration of solved forms of the dominance constraints.

Dominance constraints can be represented as *dominance graphs*. Thus, an MRS corresponds to a dominance graph. In a dominance graph originating from an MRS, nodes correspond to MRS labels h whereas (dashed drawn) dominance edges correspond to MRS qeq-constraints. For instance, the RMRS from Figure 2 which was produced by HOG is translated by UTool into dominance constraints which is shown as a dominance graph in Figure 3.

UTool computes all solved forms of the dominance graph according to the algorithm that is based on graph connectivity (see [17]). One possible reading of the dominance graph from Figure 3 (and simultaneously of the RMRS in Figure 2) is given as a solved dominance graph in Figure 4.

Consequently, from each solved dominance graph, UTool recursively generates a set of pluggings. The plugging for Figure 4 looks like the following one:

```
h24(h27,h11(h17(h20,h14),h3(h7,h8)))
```

Finally, we parse all pluggings sequentially and replace each label by its corresponding predicates according to the original MRS. After that procedure is finished, we get the formula for the first plugging:

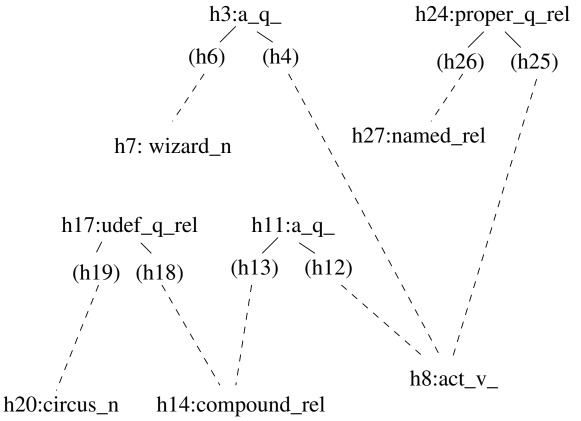

Fig. 3. Dominance graph.

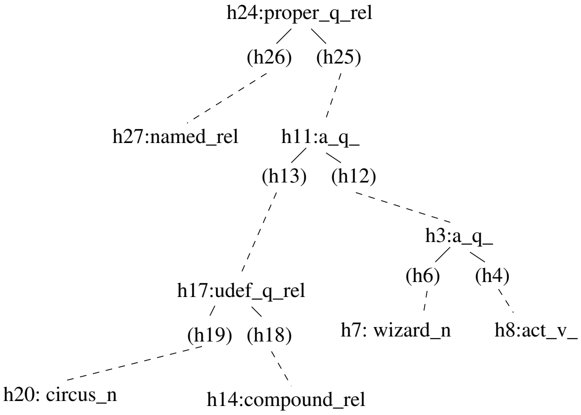

Fig. 4. Solved dominance graph.

```
proper_q_rel(X22,
    named_rel(X22,paris),
    a_q_rel(X9,
        udef_q_rel(X15,
            circus_n_1_rel(X15),and(
            compound_rel(E16,X9,X15),
            show_n_of_rel(X9))),
            a_q_rel(X5,
                wizard_n_1_rel(X5),and(
                act_v_1_rel(E2,X5),and(
                in_p_rel(E10,E2,X9),
                in_p_rel(E23,E2,X22))))))
```

IV. Applications Using FOL Expressions

There are many applications depending on a logical representation of natural language (see, e.g., [1] or [4]). The application presented here was successfully implemented in our experimental system for recognizing textual entailment (RTE) [9]. In RTE [19], the aim is to identify the logical relations between two texts, thesis T and hypothesis H, e.g.,

```
T: A wizard acts in a circus show in Paris.
H: Some magician remains in a capital
   town of France.
```

In particular, given a pair $\{T, H\}$, our system was designed to find answers to the following conjectures [20]:

1) T entails H,
2) $T \wedge H$ is inconsistent, or
3) H is informative with respect to T, i.e., is H a new and consistent information in relation to T?

In the example above, T entails H. To prove it, our system uses model-theoretic approach combined with logical inference. Syntax and semantics of the texts representing T and H are first analyzed with HOG and the resulting semantic representations in form of RMRS are translated into FOL as described in Section III. Afterwards, they are passed to external automated reasoning tools like *model builders* and *theorem provers* which check what kind of logical relation between T and H holds.

V. Conclusion and Future Work

In this paper an application based on a combination of linguistic resources and tools is presented that enable for an efficient generation of first-order logic formulas from a broad coverage HPSG grammar combined with shallow analyses. Since the semantic composition is performed by a deep HPSG parsing, it is required for a successfull generation of FOL expressions that the deep parsing succeeds. Unfortunately, it can fail if well-formedness of syntactical structures is too weak, e.g., in SMS dialogues or transcriptions of spontaneous speech.

The application can be improved through integration of coreference resolvers, so that different object variables pointing to the same individual can be identified. Finally, the application could be supplemented with statistical models describing scopal position settings in natural language sentences, so that scope resolved readings are produced in order of descending probability of their occurrence.

Furthermore, RMRS is the common semantic formalism for the HPSG grammars within the context of the *LinGO Grammar Matrix* [21] like the Japanese HPSG grammar *JaCY* [22], the *Korean Resource Grammar* [23], the *Modern Greek Resource Grammar* [24], the Norwegian *NorSource Grammar* [25], and the Spanish Resource Grammar *SRG* [26]. Because all of these grammars interface to RMRS, an exchange of the ERG in our system can be considered and a high degree of multilinguality achieved.

Acknowledgment

The authors would like to thank Matthias Hecking for many fruitful discussions on the work presented in this paper.

References

[1] P. Blackburn, J. Bos, M. Kohlhase, and H. D. Nivelle, "Inference and Computational Semantics," in *In Third International Workshop on Computational Semantics (IWCS-3).* Kluwer, 1998.

[2] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database.* MIT Press, 1998.

[3] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A Core of Semantic Knowledge," in *16th international World Wide Web conference (WWW 2007).* New York, NY, USA: ACM Press, 2007.

[4] J. Sowa, *Knowledge Representation: Logical, Philosophical and Computational Foundations.* Pacific Grove, CA: Brooks/Cole, 2000.

[5] A. Copestake and D. Flickinger, "An open-source grammar development environment and broad-coverage English grammar using HPSG," in *Proceedings of the Second conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, 2001.

[6] A. Copestake, D. Flickinger, C. Pollard, and I. A. Sag, "Minimal recursion semantics: An introduction," *Research on Language and Computation*, vol. 3, pp. 281–332, 2005.

[7] U. Schäfer, "Integrating deep and shallow natural language processing components – representations and hybrid architectures," Ph.D. dissertation, Faculty of Mathematics and Computer Science, Saarland University, 2007.

[8] P. Blackburn and J. Bos, "Underspecification, Resolution and Inference for Discourse Representation Structures," 2004.

[9] A. Wotzlaw and R. Coote, "Recognizing Textual Entailment with Deep-Shallow Semantic Analysis and Logical Inference," in *Proceedings of the 4th International Conference on Advances in Semantic Processing (SEMAPRO 2010)*, Florence, Italy, 2010.

[10] H. Bunt, "Semantic underspecification: Which technique for what purpose?" in *Computing Meaning*, H. Bunt and R. Muskens, Eds. Springer, 2007, vol. 3.

[11] A. Copestake, "Report on the design of RMRS," University of Cambridge, Tech. Rep., 2003.

[12] A. Koller and S. Thater, "Efficient solving and exploration of scope ambiguities," Proceedings of the ACL-05 Demo Session, 2005.

[13] T. Brants, "TnT - A Statistical Part-of-Speech Tagger," in *Proceedings of Eurospeech*, 2000.

[14] W. Drożdżyński, H.-U. Krieger, J. Piskorski, U. Schäfer, and F. Xu, "Shallow Processing with Unification and Typed Feature Structures — Foundations and Applications," *Künstliche Intelligenz*, vol. 1, 2004.

[15] U. Callmeier, "PET. A Platform for Experimentation with Efficient HPSG Processing Techniques," *Journal of Natural Language Engineering*, vol. 6, no. 1, 2000.

[16] A. Copestake, "An Algebra for Semantic Construction in Constraint-based Grammars," in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, 2001.

[17] S. Thater, "Minimal recursion semantics as dominance constraints: Graph-theoretic foundation and application to grammar engineering," Ph.D. dissertation, 2007.

[18] A. Koller and S. Thater, "An improved redundancy elimination algorithm for underspecified representations," in *ACL*, 2006.

[19] I. Dagan, B. Dolan, B. Magnini, and D. Roth, "Recognizing textual entailment: Rational, evaluation and approaches," *Natural Language Engineering. Special Issue on Textual Entailment*, vol. 15, no. 4, pp. i–xvii, 2009.

[20] P. Blackburn and J. Bos, *Representation and Inference for Natural Language. A First Course in Computational Semantics.* CSLI, 2005.

[21] D. F. Bender, Emily M. and S. Oepen, "The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars ," in *Procedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics.*, Taipei, Taiwan., 2002.

[22] M. Siegel and E. M. Bender, "Efficient deep processing of japanese." in *In Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization. Coling 2002 Post-Conference Workshop*, Taipei, Taiwan, 2002.

[23] K. Jong-Bok and Y. Jaehyung, "Parsing mixed constructions in a typed feature structure grammar," in *Lecture Notes in Artificial Intelligence*, vol. 3248. Springer, Feb. 2005.

[24] V. Kordoni and N. Julia, "Deep analysis of modern greek," in *Lecture Notes in Computer Science*, J.-H. L. Keh-Yih Su, Jun'ichi Tsujii, Ed., vol. 3248. Springer, Berlin 2005.

[25] L. Hellan, "From Grammar-Independent Construction Enumeration to Lexical Types in Computational Grammars," in *Coling 2008: Proceedings of the workshop on Grammar Engineering Across Frameworks.* Manchester, England: Coling 2008 Organizing Committee, August 2008, pp. 41–48.

[26] M. Montserrat, B. Núria, E. Sergio, and S. Natalia, "The spanish resource grammar: pre-processing strategy and lexical acquisition," in *Proceedings of the Workshop on Deep Linguistic Processing, Association for Computational Linguistics (ACL-DLP-2007)*, T. e. a. Baldwin, Ed., 2007.