NATURAL LANGUAGE PROCESSING USING HPSG

Milan HOLUB, Master Degree Programme (4)
Dept. of Information Systems, FIT, BUT
E-mail: xholub10@stud.fit.vutbr.cz

Supervised by: Dr. Alexander Meduna

ABSTRACT

Head-driven Phrase Structure Grammar - natural language processing theory is presented. The theory is based on assumption of the head constituent of a phrase and employs mechanism called unification which governs language analysis. Formal model called sign which bears the language information and which simplifies unification process is explained.

1 HPSG

HPSG is an abbreviation from Head-driven Phrase Structure Grammar. The basic idea of the theory is incorporated in a notion of *the head constituent of a phrase*. The theory assumes that in each phrase of any natural language there exists a main element which drives the meaning of the whole phrase.

Natural language is according to HPSG uniquely described by following equation:

$$Language = \underbrace{P_1 \wedge \cdots \wedge P_n}_{UG} \wedge \underbrace{\cdots \wedge P_{n+m}}_{LSC} \wedge (\underbrace{L_1 \vee \cdots \vee L_p}_{LS} \vee \underbrace{R_1 \vee \cdots \vee R_q}_{GR}) \quad (1)$$

The equation denotes that an object is a natural language sign(will be discussed later) token just in case it satisfies all the universal(UG) and language specific constraints(LSC), and either it instantiates one of the language lexical signs(LS) or it instantiates one of the language grammar rules(GR).

HPSG theory employs *feature structures* to describe linguistic information.

A feature structure is an information-bearing object that describes another thing by specifying values for certain kind of attributes of the described thing. Standard notation for feature structures are *attribute-value matrices* (AVM). Feature structure notated by AVM is called a *sign*. An important property of feature structures is their potential for hierarchiality.

Sign can bears either information about entry in dictionary - in this case it is called *lexical sign* - or, which is more interesting, it can hold the whole phrase and then it is called *phrasal sign*. One of the main features of sings is that we can ¨combine¨ more signs together. This process is called *unification*.

1.1 SYNTACTIC FEATURES AND SYNTACTIC CATEGORIES

We distinguish two types of (syntactic) features within the sign: local and binding. Local features in general specify inherent syntactic properties of a sign (e.g. inflection, case, lexicality), on the other hand, binding features provide non-local information about dependent elements within a sign (e.g. relative pronouns, interrogative expressions).

Among the local features we distinguish between **HEAD feature**(specifies syntactic properties that lexical sign shares with its phrasal signs; it determines the head constituent of a phrase), **SUBCAT feature**(gives information about the valence of a sign, i.e. the number and kind of phrasal signs that the sign subcategorizes for) and **LEX feature**(this is binary feature which distinguishes between lexical and phrasal sign)

Among the binding features we distinguish between SLASH, REL and QUE features. The SLASH feature provides information about gaps within a sign which have not yet been bound to an appropriate filler; the REL and QUE features give information about unbound relative and interrogative (question) elements within the sign.

Let us move to head constituent of a phrase now. This notion is incorporated into HPSG theory by Head Feature Principle.

It belongs to principles of universal grammar (it is shared among all natural languages). Simply said the principle designates following: If a phrase has a head daughter, then they share the same head features.

1.2 SOME NOTES ABOUT SEMANTICS

If we want to work with semantic meaning of utterances we need first to specify the way of classifying things in the world. HPSG uses *scheme of individuation*. This can be thought of as a system for breaking up reality into comprehensible parts which are called **individuals**(such as Alexander Meduna, planet Earth etc.), **properties**(such as being a student of BUT) and **relations**(such as loving, hating or giving).

So far to semantics - interested reader can consult Pollard [2].

1.3 SUBCATEGORIZATION

Subcategorization is one of the main and the most important concepts in HPSG. Subcategorization, sometimes called valence as well, of a lexical or phrasal sign, is a specification of the number and kind of other signs that the sign in question characteristically combines with in order to become complete. It describes dependencies that are hold between a lexical head and its complements. Attribute SUBCAT takes a list of (partially specified) signs as its value; the position of an element on the list corresponds to the obliqueness of the complement sign which it describes, with the rightmost element corresponding to the least oblique dependent sign (in case of verb it is subject). The most common use of the SUBCAT feature is in connection with verbs which usually subcategorize beside subject also for direct or sometimes for indirect object. See example below with **verb** force in past tense.

We can find 3 complements in SUBCAT list of verb **force**. The right most is the least oblique element - subject, followed by direct object and finally the 3rd element represents verb phrase.

$$[\textbf{He}]_{NP-subject}\ \textbf{forced}\ [\textbf{me}]_{NP-direct-object}\ [\textbf{to write the report}]_{VP[INF]}.$$

In accordance with SUBCAT feature we distinguish 2 types of signs: *saturated signs*(their SUBCAT list is empty - e.g. **Kim left**) and *unsaturated signs*(SUBCAT list is not empty - e.g. **forced me to write the report**).

The transmission of subcategorization within the sign is governed by a principle of universal grammar called Subcategorization Principle.

1.4 GRAMMAR RULES

Each language presents a finite set of lexical signs (vocabulary) and a finite set of grammar rules. A grammar rule is just very partially specified phrasal sign which constitutes one of the options offered by the language in question for making more concrete signs (bears more information) from less concrete ones(see also equation **(1)**).

2 CONCLUSIONS

In this paper we presented one of the approaches to natural language processing. HPSG is information based approach which meets most of linguistic requirements. We are able to analyze many aspects of language including subject-predicate relation, inflectional verbs (employing SUBCAT feature), interrogative (with help of binding features) and many more. We can strictly distinguish between syntactic and semantic properties that are valid across all languages and those that are applicable only in a particular language. This property makes the theory more versatile because of using ¨common base¨ in analysis of languages.

On the other hand, the versatility and generality of principles and rules assumes complexity of lexicon.

In my future work I would like to focus on implementation HPSG theory on the Linux platform. There are many practical applications of the theory: from grammar checkers to complex translational systems from one natural language to another.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Alexander Meduna without whose enthusiasm and professional advice writing of this paper would have been a far longer and far less pleasant task.

REFERENCES

[1] Meduna A.: Automata and Languages: Theory and Applications, London, Springer 2000, ISBN 1-85233-074-0

[2] Pollard, C. and Sag I. A.: Information-Based Syntax And Semantics, Volume 1 Fundamentals, Stanford, CSLI 1987, ISBN 0-937073-23-7