Top Ontology as a Tool for Semantic Role Tagging

Karel Pala and Pavel Smrz

Faculty of Informatics, Masaryk University
Botanicka 68a, 60200 Brno, Czech Republic
{pala,smrz}@fi.muni.cz

Abstract

This paper deals with the semantic roles in verb valency frames. It examines ontologies, particularly Top Ontology from the EuroWordnet 1,2 project in relation to the tagging of semantic roles of verb complements. We are aiming at a consistent system of semantic role tags that would form a base for lexico-semantic constraints integrated into a natural language parser. A new notation that exploits semantic roles (deep cases) based on the EuroWordNet Top Ontology and the set of the Base Concepts is also presented as well as preliminary statistical results of our research effort.

Introduction

The success of almost any realistic NLP application depends to a considerable extent on the quality of the lexical data used in it. Verbs play a crucial role in natural language sentences. Thus, the analysis of verbs (as sentence relational elements) and their complements with various semantic roles constitutes one of the most fruitful directions in lexical semantic research and the development of reliable electronic lexical resources.

Verbs are usually described by means of their valency frames. They can contain both the syntactic information about the verb construction itself (e.g., for English, what particles and/or prepositions are associated with a verb, or, for Czech and other inflective languages, what surface cases can be present on the surface level), and the semantic roles (deep cases) that are determined by the meaning of the verb (in a particular sense).

This paper deals mainly with the semantic roles in verb valency frames. We are aiming at a consistent system of semantic role tags that would form a base for lexico-semantic constraints integrated into a natural language parser. The requirements of such a system are as follows:

- The tags should offer labels for all the semantic roles postulated in the standard theories and also for some others, e.g. FrameNet (Baker et al., 1998), SALSA (Pinkal et al., 2003), ValLex (Lopatkova, Zabokrtsky, 2002), Valency Dictionary of Czech Verbs (Pala, Sevecek, 1997);
- It should cover a reasonably large number of lexical units, i. e. at least 5,000 verbs for a given language.

Moreover, the tagging should enable a more adequate sub-categorization of the roles, which are typically too general and thus do not describe the real lexical data adequately. The empirical adequacy of the existing semantic tags is the crucial requirement in our research effort.

The approach presented in this paper shares some characteristics with that of FrameNet. The main difference is that FrameNet frames are in a way more detailed then our valence frames, they go rather in the direction of Minsky's frames, with the inevitable danger of not being enough general. Our experiments with approx. 1,000 most frequent Czech verbs clearly show that our system is usable for the general verb lexicon.

Need for Subcategorization of Semantic Roles (or Deepen Deep Cases)

There are several well-established theories for verb frames and semantic roles of their participants so the question could be why not apply one of them. Let's present a simple example that will easily explain our reasons:

Take e.g. the verb vstoupit/to enter and the following two sentences:

Jan vstoupil do Komunistické strany v roce 1948.
/Jan entered the Communist party in 1948.
Jan vstoupil do budovy před 10 min.
/Jan entered the building 10 min. ago.

If we use an existing inventory of the roles then the constituents strana/party and budova/building would be most likely labeled as PAT(iens). However, the general role of patiens does not allow distinguishing different senses of the verb, e.g. Jan could be a new member of the party but not of the building. We are obviously dealing with two different senses of the verb vstoupit/to enter or more precisely with vstoupit:4/enter:3 and vstoupit:1/enter:1 if we use the standard Princeton WordNet (Fellbaum, 1998) notation. Thus vstoupit:4/enter:3 means that people typically register formally as participants or members of organizations and vstoupit:1/enter:1 denotes that people come or go into places (like buildings).

Many applications that will work with the semantic representation of natural language sentences (e.g. advanced machine translation systems) will need to distinguish between such cases. We want to express this fact by means of the semantic role tags but the label PAT definitely cannot capture the outlined sense differences. Thus we need more specific subcategorization.

A similar observation can be made for other verbs, e.g. the roles associated with verbs like eat, drink or wear certainly call for subcategorization features like FOOD, BEVERAGE or GARMENT.

The solution we are offering uses two level semantic role labels. The first level contains traditional (general) tags like AG, PAT, OBJ, INSTR LOC, etc. On the second level, we have decided to take advantage of rich WordNet hierarchical structures and to use selected literals (lexical units) occurring in the particular synsets. To be more precise, the EuroWordNet Top Ontology and the set of Base Concepts (Vossen, 1999) have been employed for tagging of semantic roles in our system. It can be seen,

Verb frame	Frequency	Sense characterization
AG (person:1) = ACT (act:2)	29	solving tasks, performing activities
AG (person:1) = OBJ (object:1)	23	manipulating with objects
AG (person:1) = PAT (person:1)	21	relations between persons
AG (person:1) = 0	15	non-personal verbs, without complement
AG (person:1) = $ (ze)	15	communication activities
AG (person:1) = SUBS (food:1)	9	verbs of eating
AG (person:1) = LOC (location:1)	8	motion verbs
AG (person:1) = ACT (job:1)	7	working
AG (person:1) = OBJ (object:1) = LOC (position:1)	7	motion with objects, positioning in space
AG (person:1) = OBJ (object:1) = OBJ (object:1)	7	combining objects
AG (person:1) = ABS (abstraction:1)	6	keeping rules
AG (person:1) = ART (garment:1)	6	verbs of dressing
AG (person:1) = EVEN (result:3)	6	making conclusions
AG (person:1) = ACT (role:1)	5	being in a position (or losing it)

Table 1: The most frequent verb frames in Czech

that through them we can access the individual lexical units when we process sentences on the respective levels (morphological, syntactic and semantic). We would like to stress the fact that the WordNet hierarchical relations capture more than 115,000 synsets (in Princeton WordNet 2.0). No other freely available resource offers such a coverage.

A Case Study of Verb Frames for the Most Frequent Czech Verbs

To demonstrate usability of our approach we have prepared a list of verb frames for approx 1,000 verbs taken from the Czech verb frequency list. The semantic roles of particular participants of valency frames have been tagged using literals from Czech Wordnet containing currently about 44,000 literals in 28,000 synsets. Note that the Czech Wordnet has been extended during our experiments to cover all the processed verbs.

Notation

To simplify processing of roles, several lexical databases exploit the simple binary model of verb-participant relations. For example, EuroWordNet notation based on binary relations defines ILRs (Internal Language Relations) (such as ROLE_AGENT – ROLE_AGENT_INVOLVED) for this purpose. On the other hand, we opt for a more complex notation, which, moreover, comprises both – surface (morphological) cases and prepositions required by Czech verbs, and the respective semantic roles. The following examples demonstrate analyzed frames for the verbs mentioned above:

jíst:1 / eat:1
kdo1*AG(person:1|animal:1)=co4*SUBSTANCE(food:1)
pít:1 / drink:1
kdo1*AG(person:1|animal:1)=co4*SUBSTANCE(beverage:1)
obléci si / put on
kdo1*AG(person:1) = co4*ART(garment:1)
= na co4*BODY(body part:1)
vyprávět:1 / tell:3
kdo1*AG(person:1) = co4*INFO(message:2),
= komu3*ADR(person:1|animal:1)

The morphological cases (seven in Czech and here nominative=1, dative=3 and accusative=4) are indicated by the forms of the respective pronouns (kdo1/who, komu3/to whom, co4/what). The semantic roles are denoted by the general labels taken from the TOP Ontology together with the subcategorizing literals from the set of Base Concepts and include the numbers of the respective senses.

Preliminary Results

The created list of verb frames has been sorted according to the deep valency frames with the aim to obtain semantically relevant verb classes. If we have a look at the obtained list we can say that our assumption has been justified with some reservations, namely: the list of 1,000 verbs is not large enough yet and there is quite a large number of the obtained groups containing less than 3 items. The obtained results are shown in Table 1.

As we have expected, the discrimination power of the frames is quite closely related to the selection of the subcategorization features. The good news is that the obtained classes are not arbitrary and can (and will) be independently confirmed by the corpus data using word sketch technique (Kilgariff et al., 2004). There is also a hypothesis that the verb classes we have arrived at might in some way correspond to Levin's verb classification (Levin, 1993).

As all our verbs are linked to their English equivalents by means of ILI (Inter-Lingual Index) defined in EuroWordNet, the frames prepared for Czech verbs can be compared to their English equivalents. It would be certainly premature to claim that all the semantic roles associated with Czech verbs strictly apply to their English counterparts, but our preliminary investigation shows almost perfect agreement in semantic roles of translation equivalents between Czech and English.

Since we participate in Balkanet Project (http://www.ceid.upatras.gr/Balkanet/), we are preparing to test whether the indicated agreement would apply also to other languages, particularly to Bulgarian, Romanian and Turkish. The deep valency frames will be transformed via ILI to the corresponding Bulgarian, Romanian and Turkish verbs and the (dis)agreement will be evaluated.

Deep Valency Frames and Parsing, Consistency Checking and Selection of the Best Analysis

There are two independent parsers for Czech implemented at the Faculty of Informatics, Masaryk University, Brno. DIS/VADIS (Mrakova, 2002) is implemented in Prolog and adopts shallow parsing strategy, while SYNT (Smrz, Horak, 2000) aims at deep robust parsing of free Czech text. Both tools can work with surface valency frames and the ability to process deep valency frames with the semantic roles is now being added to them. Moreover, the parsers are able to access data from wordnets and thus obtain the necessary information about word relations.

The results of parsing can significantly improve the process of tagging semantic roles in valency frames as they can be immediately used for consistency checking of manually prepared frames. Available parsers are able to associate a particular word in the parsed sentence with the corresponding node in a wordnet database and verify whether the role tag in the relevant frame position is the hypernym (not always direct) of the given word.

The link between the parser and the wordnet can also help in refinement of existing verb valency frames. The procedure takes again the advantage of traversing hyper-/hyponymic wordnet trees and statistically evaluates the distance between the role tags and the corresponding words from the parsed sentences. Too general as well as too specific labels denoting the semantic roles can be identified and the appropriate substitution can be automatically recommended.

More ambitious goal of our research is the (semi)automatic technique for semantic role tagging. It is obvious that the procedure could significantly speed-up the process of building valency dictionaries designed as lexicons for NLP applications. The generalization of the verb complement types is based on the data from Czech and English WordNet as well as on a separate English-Czech list of verbs based on Levin's semantic classification which contains approximately 3,500 verbs. The algorithm builds on the observation that each semantic class can be typically linked to a small number of specific semantic roles only, rarely more than five or six. The current comparison of the information contained in the manually prepared frames with automatically inferred lexico-semantic constraints shows very good match but also a lot of noise due to figurative meanings and language creativity.

Parsing can help in role tagging but, of course, the same holds in the other way around as well. The preliminary investigation of SYNT outputs shows that the subcategorization of semantic roles in verb valency frames can be integrated into the procedure of the best analysis selection (Horak, Smrz, 2002). Highly ambiguous grammars are typical for robust parsing of morphologically rich languages and thus many syntactic structures are usually offered as a result of syntactic analysis, even though only one or two are the appropriate ones in the given context. The deep verb frames allow us to obtain the relevant semantic information about the sentence constituents in the course of the syntactic analysis and to radically reduce the number of ambiguous outputs.

The importance of the detailed semantic specification does not obviously end at the level of syntactic analysis. It is crucial for any subsequent step of processing with the aim to understand meaning of a given sentence. For example, we can take advantage of the synsets and the hyper-/hyponymic trees related to the constituents in the analyzed sentence and look for the semantic relations they may have to the constituents occurring in other sentences of the given text. This is very useful for semantic processing of the texts – a possible direction is, for instance, to integrate such a procedure into the process of information extraction.

Even without an embracive NLP lexicon, just using deep valency frames, which tell us what semantic roles are expected, we can also disclose what meaning of the individual sentence constituents is present and consequently infer the meaning of the sentence as a whole.

Exploiting Derivational Relations for Checking and Extending Valency Frames

Czech, as an example of the family of Slavic languages, can be characterized by rich morphology, inflective but also derivational one. The derivational morphology can describe (apart from other morphological processes) semantic relations across different parts of speech, i.e. relations like *učit/to teach – učitel/teacher – učení/teaching – učený/educated – učenec/scholar – učiliště/training institution* etc. It can be seen that such derivations center around one stem or root and create derivational nests.

The link between valency frames and derivational nests lies in the possibility of checking consistency of the frames and their extensions based on the information about derivational relations of words that can be located in the wordnet database. If we are able to determine the relation between the verb and relevant derived nouns as well as retrieve the information about synonyms of the given words from the wordnet, we can compare the given senses and check validity and completeness of valency frames.

Take again the example of two senses of the verb vstoupit/to enter. The derived noun in Czech is vstup and the information from WordNet distinguishes the following senses:

1. vstup, vchod / entrance
   (e.g. vchod do budovy / entrance to the building)
2. vstup / joining
   (e.g. vstup do strany / joining the party)

We can check whether the verb valency frames for vstoupit cover the distinction in meaning of the morphologically related nouns.

The present version of the morphological module AJKA (Sedlacek, Smrz, 2001) used for lemmatization and morphological tagging of Czech texts is able to handle some regular derivational relations between Czech word forms. We needed to link AJKA with Czech WordNet which contains only basic word forms, i.e. nominatives of nouns and adjectives, infinitives of verbs and basic forms (not comparatives or superlatives) of adverbs. The basic items in AJKA are stems (word bases), thus the task consisted in associating the stems with the individual literals occurring in synsets.

The implemented morphological interface for the Czech WordNet database enables us to track derivational (semantic) relations and also to transfer these links to the synsets in Czech WordNet. We got an independent and relevant check of the roles in the deep valency frames and we can perform the validation employing a considerable number of Czech suffixes which cover a large part of Czech word stock. As a side effect, we can enrich Czech WordNet with the derivational nests that in fact represent subnets in a large net and in this way make Czech WordNet more suitable for NLP applications. Moreover, the derivational relations can be immediately used for a sort of inferences that are different from hyper-/hyponymic and synonymic relations and are not captured by any other language resource.

Conclusions and Future Directions

Even though the number of verbs for which the deep valency frames have been prepared is still rather limited it have already stimulated further linguistics research in the direction of natural language understanding.

We will continue our work on verb valency frames and investigate the possibilities to (semi)automatically derive valency patterns for morphologically related nouns and adjectives. We will also try to find ways how to deal with the influence of figurative meanings on the procedure for automatic derivation of role labels.

References

Baker, C.F., Fillmore, C.J., Lowe, J.B. (1998). FrameNet Project. In Proceedings of the Coling-ACL, Montreal, Canada.

Fellbaum, C. (ed.) (1998). WordNet: An Electronic Lexical Database, MIT Press.

Horak, A., Smrz, P. (2002). Best Analysis Selection in Inflectional Languages. In Proceedings of the 19th International Conference on Computational Linguistics. Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing, 2002. pp. 363-368. ISBN 1-55860-894-X.

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. To appear in Proceedings of EURALEX 2004.

Levin, B. (1993). English Verb Classes and Alternations: A Preliminary Investigation, The University of Chicago Press.

Lopatkova, M., Zabokrtsky, Z. (2002). Valency Dictionary of Czech Verbs, LREC 2002.

Mrakova, E. (2002). Partial Parser DIS/VADIS (for Czech), Ph.D. Thesis, Faculty of Informatics, Masaryk University, Brno.

Pala, K., Ševeček, P. (1997). Valencies of Czech Verbs Studia Minora Facultatis Philosophicae Universitatis Brunensis, vol. A45, Brno, pp. 41-54 (in Czech).

Pinkal, M., Erk, K., Kowalski, A., Padó, S. (2003). Building a Resource for Lexical Semantics (SALSA Project), Proceedings of the 17th International Congress of Linguists, Prague.

Sedlacek, R., Smrz, P. (2001). A New Czech Morphological Analyser AJKA, Proceedings of TSD 2001, Springer Verlag, LNAI 2166, pp.100-107.

Smrz, P., Horak, A. (2000). Large Scale Parsing of Czech. In Proceedings of Efficiency in Large-Scale Parsing Systems Workshop, COLING 2000. 1st ed. Saarbrucken : Universitat des Saarlandes. pp. 43-50. ISBN 1-55860-717-X.

Vossen, P. (ed.) (1999). EuroWordNet: A Multilingual Database with Lexical Semantic Networks for European Languages, Kluwer Academic Publishers, Dordrecht.