Towards recognizing thematic roles for verbal frames by linking two independent language resources for a parser based on the supply and demand paradigm

Balázs INDIG
(Supervisor: Gábor PRÓSZÉKY)
Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
50/a Práter street, 1083 Budapest, Hungary
indig.balazs@itk.ppke.hu

***Abstract*—Amongst the objectives of the MTA-PPKE Hungarian Natural Language Research Group is to develop a psycholinguistically motivated language processing system, which can process raw text and can build rich syntactic and semantic representation [1]. One of the key steps is to recognize and classify the verb-argument relations which is found in natural language sentences such as grammar roles and thematic roles.**
In this paper we introduce the ongoing research aiming to extend the verbal construction frame database of the MTA-PPKE parser. In our work we try to use and reuse more previously developed language resources if possible.

***Keywords*-grammar; parser; thematic-roles; language-resources**

I. Introduction

The main principle of the parser is parallelism. Many different "resource-thread" works by overriding and correcting each other. The inner-workings of these resource-threads are similar to the Categorical grammars [2]. The relations between language units are characterised by correspondence between "supplies" and "demands" of the so-called "structural-threads". In this paradigm the potential arguments in the sentence such as noun phrases form "offers", which is consist of lexical, morphological, and semantic properties, that can connect with one of the compatible "expectations" or "structural prediction"[1] of the corresponding argument frames of the verbs in the sentence [3].
In the next sections we first introduce the mechanisms of our parser that handle verbal argument frames and then we outline the the possibility of extending the argument frame database used in our parser with thematic role descriptions with the help of the VerbNET English language resource.

II. Parsing verbal argument frames

To support the parsing of verbal argument frames, we use the noun phrase and verbal argument rule database *MetaMorpho* Hungarian-English (and English-Hungarian) rule-based translation system [4] which is consists of context-free, feature-structure-based rewrite rules.
One portion of the rules has one thing in common. Every right-hand-side symbol contain lexical bounding. These rules characterises the lexicon, namely semantic and other properties of the noun, adjective and adverbial phrases which may span one or multiple words [5]. Examples can be seen in table I.

Noun phrase	Features
"házörző kutya" (watchdog)	noun, countable, animate
"tegnapelőtt" (the day before yesterday)	adverb of time
"hindi" (hindi)	noun/adjective, language

TABLE I
Examples of noun phrases, that span multiple tokens and their features. (Extracted from MetaMorpho)

The other class of the interesting cases of the MetaMorpho rules contains a lexical constraint for at least one, but not all right-hand-side symbol. While for the other constituents is bound only by means of word class, morphology and semantic constraints. These rules include the ones which are lexically bound at least at the verbal position. The argument positions can be bound only by idiomatic means. For examples see table II.

Verb pattern	Description
<somebody> jövendöl (foretell) <something><to somebody>	Only the verb is bound lexically
szó esik (talk) <about something>	There are auxiliary elements in an idiomatic construction

TABLE II
Examples of verb argument frame patterns. (Extracted from MetaMorpho)

Our parser reads the input strictly left-to-right, one token in each step. The appearance of every newly read token updates the graph that representing the relations of the already appeared tokens by ammending it. Ideally, arriving at the end of the sentence means that the representation contains the

[1] which may form pairs with offers

B. INDIG, "Towards recognizing thematic roles for verbal frames by linking two independent language resources for a parser based on the supply and demand paradigm" in *PhD Proceedings Annual Issues of the Doctoral School, Faculty of Information Technology and Bionics, Pázmány Péter Catholic University – 2015.* G. Prószéky, P. Szolgay Eds. Budapest: Pázmány University ePress, 2015, pp 159– 161.

correct parse.
Using the *Humor* morphological analyser tool [6] the parser annotates the tokens with their corresponding candidate tags and subsequently a special Part-of-Speech tagger that is similar to PurePOS [7], but only uses left-context ranks the candidates, which produces a local decision problem with probability scores without the Viterbi beam search. The N best candidates then used in parallel by the parsers main component [3] to rule out the best parse.
The identifying of the verb-argument relations during the processing of the sentence is based on the principle of *supply and demand*. At the appearance of the tokens belonging to a noun phrase we try to match new lexical rules or continue the ongoing rules which span multiple tokens by using their word class and morphological properties. If we fully recognize a possible noun phrase that span one or multiple token, we try to feed them to the empty argument frames of the verbs occurred in the sentence before. The appearance of a verbal token entails the loading of all the possible argument frames, that belongs to the verb, and the parser immediately tries to fill argument positions with the already appeared and fully recognized noun phrases.

III. Identifying Thematic roles

Additionally, beyond the recognition of verb-argument relations, the argument frames are suitable for characterizing the verb-argument relations as well[2]. The thematic role descriptions, which needed for characterizing semantic representation are only available only in the 10% of the verbal argument frame database of MeatMorpho. These are made for an independent project that were using MetaMorpho to support inspection of narrative psychological structures in historical texts [8].
However, the verbal argument frame rules have an important feature: they have two sides, Hungarian and English, and they contain the English equivalent for the Hungarian side: For every source language side (Hungarian) parser rule exists a corresponding destination side (English) generator rule, which is the translation of the given verb argument construction.
So it is possible to use freely available English language resources to complete first the English side of the rules, then transform the changes back to the Hungarian side using their property that there are in pairs. Thus connecting the rule pairs of MetaMorpho with the external resource by the correct mapping of rules. Hence creating a linked resource.
A similar resource is for example is the *Unified VerbIndex (VerbNet)*, that is the product of the *SemLink project* [9]. The VerbNet is a verb dictionary that is linked with *ProbBank* a syntactically and semantically annotated corpus and with the FrameNet semantic frame database. In this resource the English verb argument frames and their syntactic and semantic informations are gathered in good quality.
Our goal is to link the rules of MetaMorpho to this unified resource and to automatically complete the English verb argument frames of MetaMorpho with thematic roles as much as possible. After this, we translate the gained information to the Hungarian side of the Verbal argument frame rules of MetaMorpho.
It is important for this task to be as precise as possible, to ease the later manual correction to not worsen the quality of MetaMorpho as it is made purely by humans without machine help.
In the next section we describe the initial practical problems of linking MeatMorpho and VerbNet together and transferring the thematic role descriptions of VerbNet into MeatMorpho by an automatic approach.

IV. Mapping the verbal argument frames of MeatMorpho with VerbNet

During the implementation it must be taken into account, that the resources differ in many way and we must *unify* these differences to harmonize the resources. Both resources might contain errors, that might have negative influence on finding of parallel frames.
The creators of MetaMorpho were only loosely restricted by rules or conventions. Partly they worked on their own way, therefore there are no written documentation on the principles of development. During our observations, in the rule database of MetaMorpho we have found numerous typographical mistake and other errors which are due to human mistakes. Some part of these errors of course can be corrected automatically by spell checker, but the initial tests shows that these errors are often from rare words, that is unknown to the spell checker.
An other significant problem, that makes difficult to harmonize the two resources is the question of American and British English spelling. While MetaMorpho was originally intently developed for the British orthography[3], on the contrary VerbNet was made for the American English spelling.
In VerbNet in contrary to the flat list structure of MetaMorpho the verbs are grouped in classes by the similarity of their frames and each class may contain multiple frame, that corresponds to all the verbs in the class. There are even a class hierarchy, so classes may have subclasses and subclasses inherits properties from the higher classes and may specify them further. See detailed statistics in table III.

Description	Number of verbs
Verbs in VerbNet	6343
Has no frames, only mentioned in other resources	2057
Has frames, possible to link	4286
Verbs occurs in only one class	2957

TABLE III
THE CLASSIFICATION OF VERBS THAT OCCUR IN VERBNET BY THEIR DISTRIBUTION

There are multi-part or phrasal verbs, which are handled differently in the aforementioned resources. In VerbNet

[2]This includes syntactic and semantic relations too, but in this paper we focus on thematic roles

[3]Even though there are some occurrence of both spelling of the same verb in different rules, due to inconsistency.

words are connected with underscore ("_") symbol, but in MetaMorpho words are connected with space and are between apostrophes. We wanted to check how many phrasal verb is in English, so we used one of the most throughout resources available, the Princeton WordNet [10]. Details can be found in table IV.

Description	No. of verbs
Number of verbs in WordNet	7440
Number of phrasal verbs in WordNet	1410
Number of phrasal verbs in VerbNet	404
Number of verb stems, from phrasal verbs in VerbNet	223

TABLE IV
THE CLASSIFICATION OF PHRASAL VERBS THAT OCCUR IN WORDNET AND VERBNET

There are about 1 to 10 ratio between the number of rules and unique verbs in MetaMorpho as seen in table V. This is due to the idiomatic or other restrictions, which each makes separate rules for the verbs. This phenomena affects little more than the third of the rules. On the other hand, during the development of MetaMorpho it was not a goal to achieve good recall on the English side of the verbs. It was enough to keep the recall high on the Hungarian side and optimize the rules for precision. We must note, that this phenomena may cause problems later.

Description	No. of verbs
Number of verb argument frame rules	30 292
Number of unique English verb stems	3505
Number of verb stems, that are not exist in VerbNet	920
Is treated misspelled or unknown by the spell checker	143
(English) Idiomatic or other restriction in rules	10694
(Hungarian) Idiomatic or other restriction in rules	8347

TABLE V
THE CLASSIFICATION OF VERBS IN METAMORPHO

According to our experiments, the 42% of the verbs of the rules of MetaMorpho is listed in multiple classes of VerbIndex, so not just the frames, but the classes is also must be disambiguated.

V. RESULTS

The most simple subset of MetaMorpho rules can be characterized by the SUBJ V [OBJ] pattern. This means about 20 000 rules. Even without using the ontologies, we could produce 1658 unambiguous and 2908 ambiguous linking, by only checking for the corresponding verbs.
However, we see great room for improvement as the ambiguous links, as far we can see, can be disambiguated using the ontologies.

VI. FUTURE WORK

We divided features from VerbNet into three groups, the prepositions, the syntactic and the semantic restraints. The latter two needed to be harmonized between the resources. In MetaMorpho, all the restraints are stored in a homogeneous flat list, that makes nearly impossible to disambiguate the different types of restraints. In VerbNet, the COMLEX [11] formalism is used[4] and the three types of constraints are stored separately, making it fairly easy to store them in different ontologies.
The harmonizing task required to set up two ontologies, that contain the logical relations between the two featureset. For example: *Human* is a subset of *Animate*.
The logic reasoning task was made by the Racer reasoner engine [12]. The ontologies will be built manually.

VII. CONCLUSION

In this paper, we introduced the verb argument handling module of our parser, that is based on the supply and demand paradigm and examined the possibility of automatic transferring of thematic roles from one language resource to another, namely from VerbNet to MetaMorpho by linking them. We described our initial steps to find links between the described verbs, and also outlined our vision of the further development to fulfill the task of automatic linking of these resources.

ACKNOWLEDGMENT

I would like to gratefully and sincerely thank to my supervisor for his guidance, understanding, patience, and my colleges for their valuable help.

REFERENCES

[1] B. Indig, "Towards a psycholinguistically motivated performance-based parsing model," in *Proceedings of PPKE-ITK Doctoral School 2014.*
[2] G. Morrill, *Categorial grammar: Logical syntax, semantics, and processing*. Oxford University Press, 2010.
[3] B. Sass, "Egy kereslet-kínálat elvű elemző működése és a koordináció kezelésének módszere." in *Tanács Attila, Vincze Veronika (szerk.) XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2015.*
[4] G. Prószéky and L. Tihanyi, "Metamorpho: A pattern-based machine translation system," in *Proceedings of the 24th'Translating and the Computer'Conference*, 2002, pp. 19–24.
[5] K. Orosz, "Főnevek szemantikai jegyei s kódolásuk a metamorpho projektben." in *Tanács Attila, Vincze Veronika (szerk.) I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2006.*
[6] G. Prószéky and A. Novák, "Computational morphologies for small uralic languages," *Inquiries into Words, Constraints and Contexts Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday*, pp. 116–125, 2005.
[7] G. Orosz and A. Novák, "Purepos 2.0: a hybrid tool for morphological disambiguation." in *Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria, 7-13 September 2013.*
[8] O. Vincze, K. Gábor, B. Ehmann, and J. László, "Technológiai fejlesztsek a nooj pszichológiai alkalmazásában," in *Tanács Attila, Vincze Veronika (szerk.) I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2006.*
[9] E. Loper, S.-T. Yi, and M. Palmer, "Combining lexical resources: mapping between propbank and verbnet," in *Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg, the Netherlands*, 2007.
[10] C. Fellbaum, *WordNet*. Wiley Online Library, 1998.
[11] R. Grishman, C. Macleod, and A. Meyers, "Comlex syntax: Building a computational lexicon," in *Proceedings of the 15th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 1994, pp. 268–272.
[12] V. Haarslev, K. Hidde, R. Möller, and M. Wessel, "The racerpro knowledge representation and reasoning system," *Semantic Web Journal*, vol. 3, no. 3, pp. 267–277, 2012.

[4] in conjunction with WordNet categories