

Methods of Linking Linguistic Resources for Semantic Role Labeling

Balázs Indig^{1,2}, Márton Miháltz², and András Simonyi²

¹ Pázmány Péter Catholic University, Faculty of Information Technology and Bionics,
Práter u. 50/a, 1083 Budapest, Hungary

² MTA-PPKE Hungarian Language Technology Research Group,
Práter u. 50/a, 1083 Budapest, Hungary
{indig.balazs,mihaltz.marton,simonyi.andras}@itk.ppke.hu

Abstract. This paper presents the process of enriching the verb frame database of a Hungarian natural language parser to enable the assignment of semantic roles. We accomplished this by linking the parser’s verb frame database to existing linguistic resources such as VerbNet and WordNet, and automatically transferring back semantic knowledge. We developed OWL ontologies that map the various constraint description formalisms of the linked resources and employed a logical reasoning device to facilitate the linking procedure. We present results and discuss the challenges and pitfalls that arose from this undertaking. We also compare our rule-based approach with that of using a state-of-the-art English semantic role labeler pipeline for the thematic role transferring task.

Keywords: linked resources, ontology, verb argument frames

1 Introduction

Semantic role labeling (SRL) is a significant step in enabling syntactically analyzed sentences to have basic semantic information in order to make sense of the meaning, making possible further applications such as semantic search, question answering, knowledge base development etc. [1]. The goal of SRL is to classify verb arguments into specific semantic roles to allow further processing at the predicate level. This paper details the process of enriching the verb frame database of a novel, psycholinguistically motivated Hungarian natural language parsing model [2, 3] to enable the assignment of thematic roles (detailed in Section 1.1).

We describe a rule-based, ontology-driven approach to transferring semantic information by linking two verb frame databases which are different with respect to surface language forms but have a lot in common at the semantic level. We also present the results of using a state-of-the-art statistical SRL system to assign thematic roles to numerous examples which are from parallel corpora. By analyzing the results achieved by the ontology-driven approach and comparing it with that of the statistical semantic role labeler we were able to test the robustness of our method and assess its performance in real-life circumstances.

1.1 Our parsing model

ANAGRAMMA is a computational text understanding approach which does not follow the traditional parsing algorithms originated from information theory that are well-established in language technology, but uses some of the principles of human sentence processing set forth in [4]. The goal of this performance-based algorithm processes linguistic input no matter how ill-formed as long as the human parser could parse it.

We employ novel ideas, such as strict left-to-right operation. Each parsing step uses a trigram window where the first of the three tokens is processed by parallel threads, sometimes with the help of the two following tokens. The basic unit of processing is a (written) word. We may also think of the series of input words as a clock signal coordinating the work of the processing threads. The aforementioned parallel threads are overriding and correcting each other, which implement the matching of “offers” and “demands” representing different levels of linguistic knowledge, in a fashion similar to Categorical Grammars [5].

The first step is the morphological analysis, but because our focus is on syntactic processing we conceptualize this process as a black box, which provides the lemma of the input and those linguistic and non-linguistic features which serve as the basis for further processing. Some of these features called “demands”, they create threads that look for suitable features that may satisfy them, while others called “supplies”, which may satisfy the demand created by already existing or future threads [3].

For example the relationships between verbs and their arguments are detected by connecting the “offers” (lexical, morphological, and semantic properties) of potential arguments such as noun, adjective and adverbial phrases to the “demands” of verb argument positions [3]. The latter are introduced by looking up the sentence’s finite verbs in a verb argument database consisting of more than 30,000 entries, developed for a machine translation project [6].

The process of ‘caching substructures’ is also well known in psycholinguistics: in human language comprehension we call it holistic processing. Further details about the parser can be found in [7]. In our model we mimic this property of human parsing: frequently occurring structures may enter the analysis with their full internal structure already in place. Multi-word expressions (proper nouns, conversation formulae, idioms, etc.) are processed in a similar way, but they do not have internal structures but behave as if they were written in a single word.

1.2 Extending verb frame resources in our parsing model by linking

Our goal was to extend the aforementioned existing verb frame database with thematic role information to enable the assignment of semantic roles in the parser to allow further semantic processing. We accomplished this by linking the verb frame database to available external linguistic resources such as VerbNet [8] and WordNet [9], and by transferring as much semantic role information as possible. The linking was achieved by mapping the different constraint description formalisms of the source and target resources using two OWL ontologies and by employing the Racer OWL reasoner [10].

2 Related Work

Semantic role labeling was pioneered by [11]. CoNLL-2005 introduced a shared task to evaluate Semantic Role Labeling approaches [12]. [1] gives an in-depth overview. A recent work [13] boosts SRL with grammar and semantic type related features extracted with the help of a Chinese Treebank and Propbank.

There are several resources that link together structured linguistic databases for NLP applications. VerbNet, which we refer to in this paper is linked to PropBank, WordNet, FrameNet and OntoNotes Sense Groupings in the Unified Verb Index [14]. UBY is a large-scale lexical-semantic resource based on the Lexical Markup Framework (LMF) and combines various resources for English and German (WordNet, FrameNet, VerbNet, Wiktionary, OntoWiktionary) [15]. BabelNet is a multilingual encyclopedic dictionary and a semantic network which connects concepts and named entities in a very large network of semantic relations by integrating resources such as WordNet, Wikipedia, OmegaWiki, Wiktionary and Wikidata [16]. The Linked Open Data concept brings together many other different semantic and linguistic ontologies via semantic web technologies such as RDF links (e.g. [17]).

3 Resources

The verb frame database originates from the MetaMorpho Hungarian-to-English rule-based machine translation system [6], which uses deep syntactic analysis for the source language. It contains more than 30,000 verb frame patterns that represent the various possible argument configurations of over 17,000 Hungarian verbs. Each frame pattern contains a verb with lexical and morphological restrictions on it, and part-of-speech, semantic, morphological and (optionally) lexical restrictions that describe the verb’s argument slots. Some argument positions are optional (are not required to be present in the sentence for the verb frame matching to hold).

For example, the following verb frame entry for “ábrándozik” (*to dream*) describes the equivalent of the English verb frame “somebody dreams about something”: `HU.VP = SUBJ(human=YES) + TV(lex="ábrándozik") + COMPL#1(pos=N, case=DEL)`. Here, the first argument position (SUBJ, for subject) is restricted to phrases that have the **human** semantic property, while the second argument position (COMPL#1, for complement) is required to be a noun phrase in the *delative* case.

There are 27 binary semantic properties, representing semantic classes, and 54 further morphological and other grammatical features describing restrictions on the argument positions in the whole database. The verb elements of each verb frame entry are described by 6 grammatical features.

Since the verb frame database originates from a MT system, each entry describing a Hungarian verb frame also has an English translation equivalent. This English verb frame contains the English equivalent verb and argument positions equivalent to the Hungarian argument positions (and optionally more slots that

introduce new tokens that constitute the semantically equivalent VP in English). The English equivalent of the verb frame shown above for “ábrándozik” is EN.VP = SUBJ + TV(lex="dream") + COMPL#1(pre="about"). This shows, for instance, that the argument slot (COMPL#1), which is expressed by a relative case marker in Hungarian, is expressed by a prepositional phrase headed by “about” in English.

Our central idea was to use the English verb frame equivalents to link the MetaMorpho (MMO) Hungarian verb frame database to an English verb semantic resource *at the argument level* in order to transfer thematic role information. We focused on VerbNet (VN), a high-quality and broad-coverage online verb lexicon for English [8, 14]. It is organized into hierarchical verb classes extending Levin’s classes [18]. Each verb class in VN contains syntactic descriptions (syntactic frames), and selectional restrictions (such as semantic types and syntactic properties) on the arguments, whose thematic roles are also described. Continuing our example, the Hungarian verb frame entry for “ábrándozik” can be mapped to the following VN frame entry for its English translation, “dream” (which belongs to the wish-62 VN verb class):

```
NP V NP
Experiencer V Theme<-sentential>
```

By using the mapping between Hungarian MMO, English MMO and English VN arguments in the linked entries, we can infer that the thematic role of the SUBJ argument of the Hungarian verb “ábrándozik” in the above verb frame is *Experiencer*, while the other argument (COMPL#1) is a *Theme*.

In VN, in contrast to the flat list structure of MMO, verbs are grouped into classes according to the similarity of their frames, and each class may contain multiple frames that are valid for all verbs in the class. There is a class hierarchy, which means that classes may have subclasses and subclasses inherit properties from the higher classes and may specify them further. See detailed figures in Table 1.

Table 1. Verbs in VerbNet

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

There is a ratio of about 1 to 10 between the number of verb frames and unique verbs in MMO, as seen in Table 2. This is due to various idiomatic and other intricacies, which produce several different frames for the majority of verbs. This phenomena affects little more than the third of the rules. On the other hand, during the development of MMO it was not a goal to achieve good recall on the

English side of the verbs. It was enough to keep the lexical coverage high on the Hungarian side and optimize the translation equivalents for the target language for precision, which presents a problem for linking.

Table 2. Verbs in MetaMorpho

Description	No.
Number of verb frames	30 292
Number of unique English verb stems	3505
Number of verb stems that are not in VerbNet	920
Verbs treated as misspelled or unknown by the spell checker	143
Idiomatic or otherwise restricted English verb frames	10694
Idiomatic or otherwise restricted Hungarian verb frames	8347

According to our measurements, 42% of the verbs in MMO are listed in multiple classes of VN. Consequently, in addition to the VN frames, the VN classes corresponding to MMO frames also had to be disambiguated. For a brief overview of MMO verbs see Table 2.

4 Linking the Resources

We used multiple knowledge sources such as WordNet and our ontologies (see Section 4.2 for details) to ensure that Hungarian verb frame entries in the MMO database are linked precisely to those entries in VN that correspond to them both syntactically and semantically, and incorrect links are eliminated.

The employed procedure was the following. First, we took English verbs contained by the resources and filtered out those that do not appear in both of them. Using this filtered verb set we created all possible connections between frames with identical English verbs, and used this maximal mapping as our baseline. In the subsequent steps we tried to reduce the number of incorrect links by applying different constraints on the mapping in an iterative development style.

In a given MMO–VN mapping the links between specific MMO and VN entries can be categorized into 5 different types:

- (i) There might not be any linked VN entry.
- (ii) Unambiguous (one-to-one) mapping: there is only one link, which can be either
 - (iia) correct or
 - (iib) incorrect.
- (iii) Ambiguous (one-to-many) mapping: there are more than one links, and they either
 - (iia) include the correct mapping (if it exists) or
 - (iib) not (possibly because it does not exist).

Because of the different granularity and level of completeness of the two resources the baseline contained a large number of entirely unsatisfactory mappings of the types (iib) and (iiib). In particular, there were many verb frames that could be found only in one of the resources, in spite of the fact that the verb itself was present in both of them. It was part of our goal to identify these entries to ease later processing.

Before applying our constraints on the baseline mapping we further reduced the number of entries by selecting only those frames from MMO that do not have optional arguments and do not require reordering of the arguments either. These mono- and ditransitive verbs had a good coverage in the original baseline set.

To determine the real-life occurrence frequencies of various MMO verb frame types, we used the Verb Argument Browser (VAB) [19, 20], a resource derived from the 180-million word Hungarian National Corpus [21]. The VAB contains analysis of 18.3 million finite verb clauses in which the finite verb and the heads of the nominal phrases that are either arguments or modifiers of the verb are annotated. We mapped the case markings of the VAB argument nominals to MMO verb frame terminology: nominative case=SUBJ, accusative case=OBJ, other case markings or postpositons=COMPL. Using these labels we counted the occurrences of each different verb frame type in the corpus. As you can see in Table 3, the top 4 types account for 88% of all verb occurrences in the corpus. Based on this, we only considered the intransitive, mono-transitive (object or complement with non-accusative case marking) and ditransitive (object and complement) frames in the further stages.

Table 3. Verb frame type occurrences in the Hungarian National Corpus

Type	Occurrences	%
SUBJ TV OBJ	5,535,334	30.22%
SUBJ TV COMPL#1	4,501,736	24.57%
SUBJ TV OBJ COMPL#1	3,859,952	21.07%
SUBJ TV	2,465,005	13.46%
<i>(13 more types)</i>	1,957,700	10.68%
Total:	18,319,727	

On this reduced set we successively applied our different constraints and checked the differences between the mappings before and after each application. In applying and fine-tuning each constraint our goal was to filter out ambiguous and incorrect links keeping as many good connections as possible.

4.1 Filters

The first constraint that was used to filter the links in the baseline mapping required the number of arguments of the linked MMO and VN frames to be equal. This step required some conversion, because in VN prepositions are treated as

separate elements of the verb frames whereas in MMO prepositions are properties of the argument slots.

As a further constraint we checked whether the verb on the Hungarian side of the MMO entry had a similar meaning to that of the English verb on the VN side. The satisfaction of this constraint could be checked only for a small fraction of the links since the available mappings between MMO and the Hungarian WordNet, on the one hand, and the Hungarian WordNet and Princeton WordNet, on the other, are incomplete. It was also checked whether the two sides of the MMO entry correspond to the same synset in WordNet.

Restrictions on argument slots of prepositional verb phrases provided an additional constraint for filtering: the prepositional restrictions had to be identical, or at least compatible for each argument position of the linked verb frames. In contrast to MMO, which specifies concrete prepositions in its descriptions of English prepositional verb frames, VN organizes prepositions into a class hierarchy and its restrictions frequently indicate only a preposition class. In these cases only the compatibility of the two prepositional restrictions could be checked by testing whether the preposition required by the MMO entry is a member of the preposition class in the VN entry.

The last two constraints that were used for filtering the links required that the syntactic and semantic restrictions in the linked MMO and VN entries had to be compatible for all argument positions. In contrast to the constraints used for the previous filters, the formalisms in which the two resources describe these restrictions were so different and, especially in the case of semantic selectional restrictions, so complex that it became necessary to introduce explicit formal representations of their logical relations in the form of two manually created OWL ontologies, and to use an OWL reasoner to check the compatibility of the restrictions. For a brief overview of the number of verbs linked by the application of the aforementioned filters see Table 3.

4.2 The Ontologies and the Reasoner

The syntactic restriction ontology. While VN relies on a rich repertoire of more than 40 features to describe syntactic restrictions, MMO’s descriptions of English frames make use only of the attributes *clausetype* (6 possible values), *poss(essive)*, *num(ber)* and *tense* (3 possible values). The syntactic restriction ontology we have created represents all syntactic VN features and all possible syntactic MMO attribute/value combinations by OWL classes, and encodes their logical relationships by equivalence axioms of varying complexity (e.g., MMO’s *poss* and VN’s *genitive* features were simply stated to be equivalent, but VN’s *sentential* feature was expressed as a boolean combination of 7 different MMO attribute/value pairs).

The semantic restriction ontology. Both VN and MMO describe selectional restrictions on verbal argument positions in terms of boolean combinations of a small number of semantic categories that are organised into ontologies. However, the two ontologies are very different: both of them contain categories that are difficult to relate to those of the the other ontology (e.g., MMO’s *punct*

(punctuation) or VN’s *communication*), and they interpret seemingly identical categories strikingly differently (e.g., in MMO’s categorisation events can be *abstract*, while VN considers *event* and *abstract* to be disjoint categories).

In view of these differences, we decided to represent the logical relationships between the selectional categories of the two systems in a single, manually created semantic restriction ontology that contains both original ontologies, together with a number of bridging concepts and axioms. The bridging concepts are high-level concepts taken from the EuroWordNet top ontology [22], which served as a starting point for the development of the VN selectional ontology [8, 35]. They are organizational devices that help expressing logical relations between MMO and VN categories in a succinct and conceptually clear form. For instance, although both ontologies contain several functional categories such as *drink* (MMO) or *instrument* (VN), neither of them had EuroWordNet’s general *function* category. Adding this concept to the OWL ontology enabled expressing generalisations about functional categories (e.g., that they are all subcategories of VN’s *concrete* category). Since neither MMO’s nor VN’s selectional restriction ontology has a detailed documentation clarifying the intended interpretation of all categories they use, in the case of many categories bridging axioms were added on the basis of a careful analysis of their actual usage in the resources.

The ontology represents bridging concepts and selectional categories by OWL classes whose names follow a uniform naming scheme that encodes their source (VN, MMO or EuroWordNet) by suffixes. There are no named individuals or properties, and axioms are limited to stating that one of the `subClassOf`, `equivalentClass` or `disjointWith` relations holds between certain boolean combinations of classes.

The reasoner. The two restriction ontologies described so far reduced the problem of determining the compatibility of MMO and VN selectional restrictions to a reasoning problem: a pair of restrictions is compatible if and only if the restriction ontology does not imply that the corresponding (typically complex) ontology classes are disjoint. The general solution to this problem required the introduction of a reasoner software component into our system. Since the two ontologies consist only of boolean axioms, a simple propositional reasoner would have been sufficient, but because of its maturity and excellent support of the OWL format we used the open source version of the Racer OWL reasoner [10], which the system accessed via the OWLlink client-server protocol [23].

5 A parser-driven approach

MMO as a rule-based translation system includes simple example sentences for every verb frame translation rule, which are supposed to match exactly the rule they belong to. These sentences were used as regression tests since each sentence had to trigger only the rule it belonged to. We used these example sentences to obtain corresponding VN frames and thematic roles for the MMO verb frames in our gold standard data set and compared the results with our annotations.

Naturally, we had to add the actual sequence of thematic roles for the manually found MMO–VN links in the gold standard as previously it contained only VN classes and frames without that information. Those MMO frames in the gold standard that had no corresponding VN class and frame pair were manually annotated with thematic roles. Using this new gold standard data set of 400 MMO verb frames and the corresponding thematic roles we were ready to measure the results obtained by using a state-of-the-art English semantic role labeler.

First, we translated the Hungarian example sentences with MMO to English. This was an important step since other translation systems would most probably have produced English predicates different from the desired ones, which were exactly the corresponding verbs in the MMO frame database. Having obtained the English versions of the Hungarian example sentences, we ran an SRL system on them that was capable of identifying predicates and labeling their arguments with semantic roles. Based on its performance and availability we chose the state of the art PathLSTM semantic role labeler [24], which utilizes lexicalized dependency path embeddings and certain binary features to identify and label semantic arguments. For tokenization, dependency parsing, and semantic predicate identification and disambiguation we used the pipeline described in the documentation of the PathLSTM source code [25], which consists of the Stanford CoreNLP WSJ tokenizer [26], the Bohnet dependency parser [27], and the mate-tools semantic role labeler [28]. PathLSTM was run with a model supporting PropBank role labeling and the resulting labels were transformed into VN thematic roles via the SemLink project’s PropBank–VN mapping.[14]³

We took only the main predicates into account that matched the verb on the English side of the corresponding MMO rule. The other identified predicates were excluded. As the used PropBank–VN SemLink mapping did not always produce unique and fully matching VN frames for the identified PropBank predicates and arguments we introduced the following rules for dealing with frame ambiguity and partial matches: For each VN frame corresponding in SemLink to a parsed PropBank predicate, if the frame had an element that did not occur in the parse then it was considered a partial match, else a full match. If there were full matches for a predicate then we dropped the partial matches and selected the element with the broadest coverage. We did the same when there were only partial matches available. We preferred those partial matches where the VN frame had fewer arguments than in the parse and the other cases were considered only after them. Relying on these rules we could assign the best matching VN frame and thematic roles to each sentence.

6 Results

6.1 Filtering

To measure the performance of our system we created a random sample of 400 MMO entries from the output of the last filter. Ambiguous entries (with a one-

³ The whole system with pretrained models can be downloaded at <https://github.com/microth/PathLSTM>.

Table 4. The number of links after subsequent filters

Description	No. of linked entries (unambiguous/ ambiguous)
Baseline set	431 / 26,560
Possible reordering needed	291 / 12,664
The lengths of MMO Hungarian and English sides are not equal	285 / 12,347
Mono- and ditransitive constructions	267 / 10,146
Equal no. of arguments both in MMO and VN	2301 / 7,745
WordNet mapping	2181 / 6,858
Prepositional restrictions	2929 / 4,610
Ontology (semantic restrs)	2967 / 4,455
Ontology (both)	2733 / 3,286

Table 5. Precision and number of links after subsequent filters with regard to the gold standard

Description	No. of linked entries (unambiguous/ ambiguous)
Baseline set	100% (9) / 98.38% (183)
Possible reordering needed	100% (9) / 98.38% (183)
The lengths of MMO Hungarian and English sides are not equal	100% (9) / 98.38% (183)
Mono- and ditransitive constructions	100% (9) / 98.38% (183)
Equal no. of arguments both in MMO and VN	100% (114) / 96.29% (78)
WordNet mapping	100% (101) / 97.14% (68)
Prepositional restrictions	90.43% (104) / 79.62% (43)
Ontology (semantic restrs)	90.98% (111) / 76.59% (36)
Ontology (both)	92.59% (100) / 70.83% (17)

to-many mapping in the output) and unambiguous ones (with a one-to-one mapping) were treated equally. The sample was processed by two independent annotators and unified by a third one. The sample contained 90 MMO entries that had no corresponding entry in VN. These entries were removed and the remaining entries together with their manually determined VN links constituted our gold standard.

Since the gold standard was not representative of the whole MMO database and we considered only those entries from each test set that were in the gold standard, only the precision of the results could be assessed reliably. We checked each filter’s output in the following way: if an MMO entry was unambiguously mapped and the mapped VN entry was identical to the one specified by the gold standard then it was considered correct, otherwise it was incorrect. In the ambiguous case set containment was used instead of equality: if the correct VN entry was in the set of linked entries then the mapping was considered correct, otherwise it was incorrect.

As can be seen in Table 4, the final mapping that was produced by our procedure contained four times more unambiguous links than the baseline, while the number of ambiguous links was radically reduced. The figures in table 5 show that the precision of the filters described in Section 4.1 was nearly perfect in the case of those unambiguously mapped MMO entries for which the gold standard specified a valid corresponding VN entry. As for ambiguous mappings, they were regarded correct if the right entry was among the linked entries, but these numbers could be weighted by the number of links, which would lead to lower values.

6.2 Parser-driven approach

We used label-based and sentence-based evaluation (see Table 6), and only the precision of the parses was considered. In total 429 sentences were parsed but only 327 sentences had at least one argument with a thematic role left after checking the frame consistency checking phase.

Table 6. Result of the parser based thematic role labeling task

	Good	All	Precision (%)
No. of Labels	428	602	71.096
No. of Frames	193	327	59.021

The gold standard data set contained mainly simple verb frames where one can easily translate arguments from English to Hungarian as no argument re-ordering is needed. In the case of the few examples where the arguments were reordered during translation we compared the automatic result to the thematic roles of the English language sentences, as it is a trivial task to reorder the arguments for specific rules in the translation system ensuring that the identified thematic roles match the correct Hungarian arguments.

7 Discussion

A number of issues made the linking of MMO and VN entries more than a trivial exercise. Some of these obstacles arose from inherent problems in the used resources.

On the one hand, the MMO verb frame database was not conceived as a general-purpose resource for NLP applications, but rather to support a specific MT system. As a consequence, the lexical coverage of verbs in the English side is low, compensated by paraphrase-like translations which are hard to look up in a lexical resource such as VerbNet. The English MMO verb frames also include a large number of idioms or semi-compositional structures (one or more of the arguments are bound lexically, eg. *take part in sg.*, *make room for sg.* etc.), which are totally absent from VerbNet. Furthermore, while the features used for

specifying selectional restrictions in the Hungarian verb frames fare well within the original MT system, the lack of a strict and formal system presents challenges when mapping to another feature system.

On the other hand, VerbNet has recursive, complex selectional restriction feature expressions, which are hard to process (4.2). Even though VN is an elaborate resource, the semantic features and categories used in the syntactic frames are not well documented, or come from vaguely documented resources, which sometimes makes their interpretation difficult or a work of guessing. We found VN to be sometimes incomplete, for example, the only intransitive frame for “knock” (class `sound_emission-43.2`) marks the subject *Theme*, while we believe a frame with an *Agent* subject exists in English (“Somebody knocked.”).

Finally, WordNet presents some problems of its own. Its noun hypernym hierarchy, which is very useful as a taxonomic network, represents a level of granularity which does not reflect general (domain-independent) language use (e.g., the immediate superclasses of “dog” cover its biological taxonomy), making graph distance-based inferences difficult. The differences between the data formats of various WordNet resources (Hungarian WordNet and different Princeton WordNet versions) also presented difficulties.

From the parser-driven approach we expected better results, but it turned out that the highly advanced statistical generalizations on which the semantic role labeler relies do not play well with the hand-crafted, linguistically motivated MMO resource we were experimenting with. The parsing results were highly inconsistent and many of the problems could have been fixed inside the parser. For example some inflected verbs resulted in non-existent PropBank classes, due to bad lemmatization. There were many cases in which the resulting predicates had nothing in common with the expected classes as some arguments were missing or some extra arguments were mistakenly detected. If a known verb is found then it would probably be better to choose from the existing frame patterns instead of trying to generalize them, as further processing usually relies on the completeness of the underlying resource. Due to this erroneous behavior the results obtained using the parser fell short of what could be expected from a highly advanced statistical parsing method. Consequently, we can draw the conclusion that currently our proposed rule-based method for the cross-language transfer of thematic roles yields better results than the parser-based alternative we described, although we expect a slight deterioration in the results if a larger number of possibly more complex examples is compared to an extended gold standard.

8 Conclusion

In this paper, we presented the verb frame database that is used in our Hungarian natural language parsing model, and our initiative to link it to the VerbNet English verb lexicon, by exploiting the available English verb frame translations. The goal was to transfer the thematic role information available in VerbNet to Hungarian verb frames. We created two ontologies to harmonize the different de-

scriptive formalisms of the two resources, and applied a logic reasoner to disambiguate candidate links based on translations. While this methodology presents some issues and does not present a full-fledged solution, it enabled us to enrich our verb database with thematic role information in a way that did not require the costly manual processing of all resources.

We also experimented with a parser-driven approach that acquires the thematic roles from translated sentences, but this method utterly failed compared to the rule-based approach on a moderate sized gold standard data set because of the inconsistencies between the parser and the lexical resources. As more and more components come into play, the issue of inconsistency between the components assumes a major role that cancels the positive effects and yields worse results than fewer but consistent components and a more rigid rule-based approach.

9 Acknowledgements

An earlier, shorter version of this paper, in which the evaluation was based on a substantially smaller gold standard and a smaller set of frames (excluding complements), was presented at the 7th Language & Technology Conference in Pozna in 2015 [29]. Another paper detailing the publicly available underlying ontology was presented at LREC 2016 in Protorož [30].

References

1. Palmer, M., Gildea, D., Xue, N.: Semantic role labeling. *Synthesis Lectures on Human Language Technologies* **3**(1) (2010) 1–103
2. Prószéky, G., Indig, B., Miháltz, M., Sass, B.: Egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozási modell felé. In: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2014. (2014)
3. Sass, B.: Egy kereslet-kínálat elvű elemző működése és a koordináció kezelésének módszere. In: XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2015. (2015)
4. Pléh, C.: Formal connexity and pragmatic cohesion in anaphora interpretation. In: *Text and discourse connectedness. Proceedings of the Conference on Connexity and Coherance*, Urbino. (1989) 137–52
5. Morrill, G.: *Categorial grammar: Logical syntax, semantics, and processing*. Oxford University Press (2010)
6. Prószéky, G., Tihanyi, L.: MetaMorpho: A pattern-based machine translation system. In: *Proceedings of the 24th Translating and the Computer Conference*. (2002) 19–24
7. Prószéky, G., Indig, B.: Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel. *Alkalmazott Nyelvtudomány* **XV**(1–2) (2015) 29–44
8. Schuler, K.K.: *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania (2005)
9. Fellbaum, C.: *WordNet: An electronic lexical database*. MIT Press (1998)
10. Haarslev, V., Hidde, K., Möller, R., Wessel, M.: The RacerPro knowledge representation and reasoning system. *Semantic Web Journal* **3**(3) (2012) 267–277

11. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. *Comput. Linguist.* **28**(3) (September 2002) 245–288
12. Carreras, X., Màrquez, L.: Introduction to the CoNLL-2005 shared task: Semantic role labeling. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning. CoNLL 2005*, Stroudsburg, PA, USA, ACL (2005) 152–164
13. Ku, L.W., Virk, S.M., Lee, Y.H.: A dual-layer semantic role labeling system. *ACL-IJCNLP 2015* (2015) 49
14. Loper, E., Yi, S.T., Palmer, M.: Combining lexical resources: mapping between PropBank and VerbNet. In: *Proceedings of the 7th International Workshop on Computational Linguistics*, Tilburg. (2007)
15. Gurevych, I., Eckle-Köhler, J., Hartmann, S., Matuschek, M., Meyer, C.M., Wirth, C.: UBY – A large-scale unified lexical-semantic resource based on LMF. In: *Proceedings of the 13th Conference of the European Chapter of the ACL (EACL 2012)*. (apr 2012) 580–590
16. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* **193** (2012) 217–250
17. Schmachtenberg, M., Bizer, C., Jentzsch, A., Cyganiak, R.: Linking open data cloud diagram 2014. <http://lod-cloud.net/> (2014)
18. Levin, B.: *English verb classes and alternations: A preliminary investigation*. University of Chicago Press (1993)
19. Sass, B.: A unified method for extracting simple and multiword verbs with valence information. In: *Proceedings of RANLP 2009*, Borovec, Bulgaria. (2009) 399–403
20. Sass, B.: The verb argument browser. In Sojka, P., Horák, A., Kopeček, I., Pala, K., eds.: *Text, Speech and Dialogue, 11th International Conference, TSD 2008*, Brno, Czech Republic, September 8–12, 2008. *Proceedings*. Volume 5246 of *Lecture Notes in Computer Science.*, Springer (2008) 187–192
21. Váradi, T.: The Hungarian National Corpus. In: *Proceedings of the Second International Conference on Language Resources and Evaluation*, Las Palmas. (2002) 385–389
22. Vossen, P., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., Roventini, A., Bertagna, F., Alonge, A., Peters, W.: *The EuroWordNet base concepts and top ontology*. Technical report, EuroWordNet project (1998)
23. Liebig, T., Luther, M., Noppens, O., Wessel, M.: OWLlink. *Semantic Web – Interoperability, Usability, Applicability* **2**(1) (2011) 23–32
24. Roth, M., Lapata, M.: Neural semantic role labeling with dependency path embeddings. In: *Proceedings of ACL 2016*, Berlin. (2016) 1192–1202
25. Roth, M., Lapata, M.: PathLSTM. <https://github.com/microth/PathLSTM> (2017) GitHub repository.
26. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: *Proceedings of ACL 2014: System Demonstrations*. (2014) 55–60
27. Bohnet, B.: Very high accuracy and fast dependency parsing is not a contradiction. In: *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics (2010) 89–97
28. Björkelund, A., Hafdell, L., Nugues, P.: Multilingual semantic role labeling. In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, Association for Computational Linguistics (2009) 43–48
29. Indig, B., Miháltz, M., Simonyi, A.: Exploiting linked linguistic resources for semantic role labeling. In: *7th Language and Technology Conference: Human Lan-*

- guage Technologies as a Challenge for Computer Science and Linguistics, Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu (2015) 140–144
30. Indig, B., Miháلتz, M., Simonyi, A.: Mapping ontologies using ontologies: Cross-lingual semantic role information transfer. In Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (may 2016)