Exploiting Linked Linguistic Resources for Semantic Role Labeling

Balázs Indig[1, 2], Márton Miháltz[1, 2], András Simonyi[1, 2, 3]

[1]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
[2]MTA-PPKE Hungarian Language Technology Research Group
Práter u. 50/a, 1083 Budapest, Hungary
[3]Analogy Zrt.
Logodi u. 44, 1012 Budapest, Hungary
{indig.balazs, mihaltz.marton}@itk.ppke.hu, andras.simonyi@analogy.co

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.

1. Introduction

Semantic role labeling (SRL) is a significant step in making sense of the meaning of natural language sentences, enabling further applications such as semantic search, question answering, knowledge base development etc. (Palmer et al., 2010). The goal of SRL is to identify the semantic arguments associated with the verbs of a sentence and their classification into specific (thematic) roles. This paper details the process of enriching the verb frame database of a novel, psycholinguistically motivated Hungarian natural language parser (Prószéky et al., 2014; Sass, 2015) to enable the assignment of thematic roles.

This parser employs novel ideas, such as strict left-to-right operation, parallel threads overriding and correcting each other, which implement the matching of "offers" and "demands" representing different levels of linguistic knowledge, in a fashion similar to Categorial Grammars (Morrill, 2010). 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 (Sass, 2015). 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 (Prószéky and Tihanyi, 2002).

Our goal was to extend this existing verb frame database with thematic role information to enable the assignment of semantic roles in the parser. We accomplished this by linking the verb frame database to available external linguistic resources such as VerbNet (Schuler, 2005) and WordNet (Fellbaum, 1998), 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 (Haarslev et al., 2012).

The rest of this paper is organized as follows: Section 2. describes related work. Section 3. gives detailed account of the Hungarian-English verb frame database and VerbNet, which are linked together to exploit information. Section 4. details the process used to accomplish this linking, including our mapping ontologies and the reasoning process. Section 5. offers an overview of our results, while Section 6. discusses the outcome and pitfalls of the process. Finally, Section 7. presents possibilities for future work and Section 8. draws our conclusions.

2. Related Work

Semantic role labeling was pioneered by (Gildea and Jurafsky, 2002). CoNLL-2005 introduced a shared task to evaluate Semantic Role Labeling approaches (Carreras and Màrquez, 2005). (Palmer et al., 2010) gives an in-depth overview. A recent work (Ku et al., 2015) 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 (Loper et al., 2007). 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) (Gurevych et al., 2012). 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 (Navigli and Ponzetto, 2012). The Linked Open Data concept brings together many other different semantic and linguistic ontologies via semantic web technologies such as RDF links (e.g. (Schmachtenberg et al., 2014)).

3. Resources

The verb frame database originates from the MetaMorpho Hungarian-to-English rule-based machine translation system (Prószéky and Tihanyi, 2002), 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).

As an 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(prep="about")`. This shows, for instance, that the argument slot (`COMPL#1`), which is expressed by a delative 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 (Schuler, 2005; Loper et al., 2007). It is organized into hierarchical verb classes extending Levin's classes (Levin, 1993). 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.

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

Table 1: Verbs in VerbNet

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.

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

Table 2: Verbs in MetaMorpho

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 VN entry linked to the MMO entry in question; (ii) one-to-one, which can be either (iia) correct or (iib) incorrect; or (iii) one-to-many, which may (iiia) contain the correct mapping (if it exists) or (iiib) 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. 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 (Vossen et al., 1998), which served as a starting point for the development of the VN selectional ontology (Schuler, 2005, 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 (Haarslev et al., 2012), which the system accessed via the OWLlink client-server protocol (Liebig et al., 2011).

5. Results

Description	No. of linked entries (unambiguous/ ambiguous)
Baseline set	429 / 26552
Possible reordering needed	201 / 8603
The lengths of MMO Hungarian and English sides are not equal	183 / 4844
Mono- and ditransitive constructions	165 / 4844
Equal no. of arguments both in MMO and VN	1526 / 3259
WordNet mapping	1450 / 2837
Prepositional restrictions	1593 / 2607
Ontology (semantic restrs)	1649 / 2512
Ontology (both)	1654 / 2116

Table 3: The number of links after subsequent filters

To measure the performance of our system we created a random sample of 100 MMO entries from the output of the last filter. Ambiguous entries (with a one-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 12 MMO entries that had no corresponding entry in VN. These entries were removed and the remaining 88 MMO entries together with their manually determined VN links constituted our gold standard.

Description	No. of linked entries (unambiguous/ ambiguous)
Baseline set	100% (4) / 97.56% (80)
Possible reordering needed	100% (4) / 97.56% (80)
The lengths of MMO Hungarian and English sides are not equal	100% (4) / 97.56% (80)
Mono- and ditransitive constructions	100% (4) / 97.56% (80)
Equal no. of arguments both in MMO and VN	100% (65) / 90.47% (19)
WordNet mapping	100% (56) / 92.30% (12)
Prepositional restrictions	100% (58) / 90.90% (10)
Ontology (semantic restrs)	100% (62) / 85.71% (6)
Ontology (both)	98.38% (61) / 83.33% (5)

Table 4: Precision and number of links after subsequent filters with regard to the 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 3, 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 4 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. 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-compositonal 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.

7. Future Work

We expect manual corrections to each aforementioned resources driven by the conclusions we reported. This may aid further linking. We would also find helpful to establish some direct connections between parsers and MMO-VN to directly classify raw text from parallel corpora in order to have an example-based linkage between the two resources. Finally, we are planning to extend the reduced baseline set by allowing links between more complex entries with optional elements and by allowing argument reordering.

8. Conclusion

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

9. References

Carreras, Xavier and Lluís Màrquez, 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *CoNLL 2005*.

Fellbaum, Christiane (ed.), 1998. *WordNet*. MIT Press.

Gildea, Daniel and Daniel Jurafsky, 2002. Automatic labeling of semantic roles. *Comput. Linguist.*, 28(3):245–288.

Gurevych, Iryna, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth, 2012. UBY – A large-scale unified lexical-semantic resource based on LMF. In *EACL 2012*.

Haarslev, Volker, Kay Hidde, Ralf Möller, and Michael Wessel, 2012. The RacerPro knowledge representation and reasoning system. *Sem. Web Journal*, 3(3):267–277.

Ku, Lun-Wei, Shafqat Mumtaz Virk, and Yann-Huei Lee, 2015. A dual-layer semantic role labeling system. *ACL-IJCNLP 2015*:49.

Levin, Beth, 1993. *English verb classes and alternations: A preliminary investigation*. U. of Chicago Press.

Liebig, Thorsten, Marko Luther, Olaf Noppens, and Michael Wessel, 2011. OWLlink. *Semantic Web – Interoperability, Usability, Applicability*, 2(1):23–32.

Loper, Edward, Szu-Ting Yi, and Martha Palmer, 2007. Combining lexical resources: mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg*.

Morrill, Glyn, 2010. *Categorial grammar: Logical syntax, semantics, and processing*. Oxford University Press.

Navigli, Roberto and Simone Paolo Ponzetto, 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Palmer, Martha, Daniel Gildea, and Nianwen Xue, 2010. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Prószéky, Gábor, Balázs Indig, Márton Miháltz, and Bálint Sass, 2014. Egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozási modell felé. In *XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2014*.

Prószéky, Gábor and László Tihanyi, 2002. Metamorpho: A pattern-based machine translation system. In *Proceedings of the 24th Translating and the Computer Conference*.

Sass, Bálint, 2015. Egy kereslet-kínálat elvű elemző működése és a koordináció kezelésének módszere. In *XI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2015*.

Schmachtenberg, Max, Christian Bizer, Anja Jentzsch, and Richard Cyganiak, 2014. Linking open data cloud diagram 2014. `http://lod-cloud.net/`.

Schuler, Karin Kipper, 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis.

Vossen, Piek, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, and Wim Peters, 1998. The EuroWordNet base concepts and top ontology. Technical report, EuroWordNet project.