What Do We Drink? Automatically Extending Hungarian WordNet With Selectional Preference Relations

Márton Miháltz[1], Bálint Sass[1], Balázs Indig[2]
[1]MTA-PPKE Hungarian Language Technology Research Group, Budapest, Hungary
[2]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics, Budapest, Hungary
{mihaltz.marton,sass.balint,indig.balazs}@itk.ppke.hu

Abstract

In this paper, we describe an ongoing experiment which aims to extend Hungarian WordNet with new verb-noun relations that specify selectional restrictions for various argument positions. We present an algorithm that uses frequency data from a representative corpus and information from a verb frame description database to generate sets of semantic classes, represented by WN hypernym sub-networks. The method intends to cover all possible argument positions of verbs found in the corpus which are marked by various case inflections or postposition particles. The new links in HuWN are assigned corpus-based probabilities. We present some preliminary results and discuss some of the arising issues.

1 Introduction

Since its first release in 1985, Princeton WordNet (PWN) (Fellbaum, 1998) has become a de facto standard lexical semantic resource for natural language processing research and applications. Its availability, vast lexical coverage and solid development over the years helped it achieve a prominent status.

Over its history, a number of possibilities for improvement of WN have become evident. From the NLP user's perspective, one of PWN's weaknesses lies in the low number of cross-part-of-speech semantic relationships it defines. Most of the existing relations across the different sub-networks for nouns, verbs, adjectives and adverbs are morphological (derivational) connections, e.g. research (verb)-researcher (noun), engage (verb)-engagement (noun) etc.

In this paper, we describe an ongoing experiment whose goal is to automatically extend Hungarian WordNet with verb-noun relations that reflect selectional preferences observed in a representative corpus. We try to automatically generalize classes of concepts (hyponym sub-graphs) that represent typical arguments for certain syntactic verb-noun relations, e.g. {to eat}-{food}, {to write}-{written material} etc. This information will be used, for example, in a project that aims to construct a novel parser for Hungarian that will in part rely on deep semantic processing of the input (Prószéky 2013).

Hungarian WordNet (HuWN) (Miháltz et al., 2008) follows the principles underlying the EuroWordNet and BalkaNet projects (Vossen, 1999, Tufiş et al., 2004). It uses Princeton Wordnet (version 2.0) as its inter-lingual index, meaning that the majority of Hungarian synsets are mapped to English WN synsets. HuWN contains localizations of the Balkanet Core Set synsets, plus additional concepts totaling 42,000 synsets. In addition to the standard semantic relations found in PWN it introduces new relations to reflect some intrinsic properties of Hungarian (Kuti et al., 2008).

The rest of this paper is organized as follows: in the next section, we briefly cover some points about verb argument syntax and semantics and present our goals. Section 3 presents related work, which is followed by the description of our proposed algorithm and the presentation of some preliminary results. The paper ends with a discussion of further work and our conclusions.

2 Background

In Hungarian, the syntactic roles of verb arguments (complements) are reflected by any of 18-34 different morphological case markings (exact number depending on the chosen linguistic theory) or by various postposition particles. Different verbs have different argument structures

which impose different morphosyntactic constraints on their arguments. These in turn correspond to different semantic types of nominal concepts: *figyel valamire* (to pay attention to something[case=SUBL]), *elkezdődik valami* (something[case=NOM] begins), *odaéget valamit* (to burn something[case=ACC]), *érdeklődik valami után* (to show interest in something[postp='after']) etc.

Connections between verbs and their nominal arguments show a range of types. On the one extreme, there are idiomatic, non-compositional verb-argument relationships where a certain sense of a verb only accepts a specific lexical element in a certain argument position, e.g.: ***hangot*** *ad valaminek* ("to give **voice**[case=ACC] to something": express one's opinion about sg), *issza a* ***szavát*** ("to drink someonone's **words**": to listen closely to someone), ***tenyerén*** *hordoz* ("to carry someone around on the **palm** of one's hand": to pamper someone) etc. On the other extreme, there are verbs that impose semantic selectional restrictions on their preferred arguments. These arguments belong to (one or more) specific semantic classes: *to eat something (food), to write something (piece of writing), to spill something (liquid)* etc. These semantic classes can productively predict which lexical items these verbs will prefer in given argument positions.

The goal of the project described in this paper is to find automatic methods in order to extend Hungarian WordNet with instances of a new type of semantic relation that links verb synsets with their typical nominal argument classes. Each of these new relation instances will have two associated properties: morphosyntactic information (the case mark or postposition) identifying the given argument position, and the strength of the connection, expressed as a probability estimated from the corpus based on the frequency of usage. For instance, the connection *{to drink}–[case=acc, p=.87]–{liquid}* designates that the arguments of the verb *drink* carrying an accusative case mark (direct object position) will fall into the semantic class represented by *{liquid}* with 87% probability (as observed in the corpus.) The synset *{liquid}* here represents itself and all its direct and indirect hyponyms, thus it also represents a class of related concepts.

3 Related Work

Charting selectional preferences is a key step in the semantic processing of written language. It involves determining which word meanings are frequent and/or allowed in a specific syntactic context of another given word. Following work by Resnik (1996, 1998), several studies relied on WordNet in the detection of selectional preferences (Clark and Weir, 2002, Ye, 2004, Calvo et al., 2005).

While recent approaches have focused on Latent Dirichlet Allocation (LDA) methods (Ritter et al., 2010, Guo and Diab, 2013, Rink and Harabagiu, 2013), we present an approach that more closely resembles Resnik (1998). It is applied to resources in Hungarian, which has not been researched previously before. Our work does not only focus on the classic problem of verb-direct object selectional preferences but all possible syntactic types of arguments (20+ in Hungarian) are considered, as recommended by Brockmann and Lapata (2003).

In contrast to approaches that only aim to define which set of words are preferred as arguments of given verbs (e.g. Erk, 2007, Tian et al., 2013, Rink and Harabagiu, 2013), in line of the approach outlined by Resnik (1998) and also adapted by Guo and Diab (2013), our research attempts to assign semantic class labels to verb argument positions, which define selectional preferences. This enables us to accomplish our goal, extending Hungarian WordNet with a new type of verb-noun (verb-argument) relation.

4 Methods

We propose an algorithm which takes a set of words (frequency list of nouns in a certain argument position of a given verb from a representative corpus) and returns a weighted list of WordNet synsets that represent them (semantic classes/generalizations representing the argument position). The resulting synsets (and the hyponym sub-graphs that they represent) should satisfy the following conditions as much as possible:

Coverage: the synset and its hyponym descendants should contain as much input corpus words as possible.

Density: the hyponym sub-graph should cover as few words as possible which were not included in the input word list.

Meaningful generalizations: the output synset and its hyponym sub-graph should express a generalization of the meanings of the corpus words in the verb argument position, but it should not be too generous. For e.g. assigning *{entity}* to all verb arguments has little or no ben-

efit as it does not give insights to the semantic preference characteristic of different verbs.

Automatic word sense disambiguation: if a word associated to a verb as an argument in the corpus has several meanings in WN, we expect the algorithm to yield relations that link the verb only to the sense(s) relevant for that argument position.

Our algorithm works as follows:

1. First, it generates all possible paths from all WN synsets that contain the input words to the root nodes in the hypernym hierarchies. All the synsets at all points in all these paths are considered as representatives of candidate semantic classes.

2. This is followed by filtering of the candidates: eliminate those candidate synsets that represent only a single corpus word and which are (1 or more degree) hypernyms of the synset that contain the word. This step is applied to omit some of the candidates that present no generalization information.

3. Next, the algorithm scores the remaining candidates based on two factors: *coverage* (how many input words they cover) and *density* (number of synsets representing input words covered by the sub-graph of a candidate to the total size of the sub-graph.) The following formula is used to calculate the score for candidate synset *c* (where *subgr(c)* is the hyponym subgraph starting from synset *c* and I_c is the subset of all input words that are covered by *subgr(c)*):

$$Score(c) = |I_c| \times \frac{|\{s \in subgr(c) : w \in s, w \in I_c\}|}{|subgr(c)|}$$

4. The top N candidate synsets are returned based on the ranking. To ensure disambiguation of input words with respect to the verb argument position, the following procedure is applied: if there are any 2 candidate synsets in the list that each contain different senses of the same input word, then the lower-ranked candidate is eliminated and the N+1. ranked candidate is added to the list. This is repeated until there are no more ambiguities.

New verb-noun relations can be added to the WN network in which the verb argument positions are semantic classes represented by the winning candidates. Link probabilities are calculated using the corpus frequencies of the input words covered by the classes (see Section 6.)

We used the database of the *Verb Argument Browser (VAB)* project (Sass, 2008), which was constructed from the 187 million-word Hungarian National Corpus (Váradi, 2002). In VAB, a simple rule-based parser was used to identify clauses, finite verbs and noun phrases (heads and their morphosyntactic properties: cases and postpositions) in all sentences of the corpus. From this, for each verb in the corpus, we extracted frequency lists of all the nouns it co-occurred with, grouped by different case markings and postpositions.

To determine the possible argument structures of each verb in the corpus (number of arguments and their morphosyntactic constraints), we relied on the lexical database of the *MetaMorpho* Hungarian-English machine translation system's syntactic parser (Prószéky et al., 2004). It contains 33,000 verb frame descriptions (argument structures for various senses) for more than 18,000 Hungarian verbs. During the construction of Hungarian Wordnet, verb synsets were linked to the corresponding verb frame descriptions in this database (Miháltz et al., 2008). This information can be used to unambiguously determine the verb synsets that will participate in the newly generated selectional preference relations.

We used a subset of the MetaMorpho syntactic analyzer's rules to identify verb argument structures in the 20.24 million sentence clauses that constitute the basis of the Verb Argument Browser database. This was done to refine the contents of the VAB database, because 1) it employed a less sophisticated parser, 2) it does not differentiate between verb complements and optional modifiers (adjuncts). By using the parser, we were able to focus on the true complements. We obtained 32,000 different verb argument frequency lists for 25,500 different verb frames to run our selectional preference class identification algorithm on.

5 Results and Discussion

Since we are still working on an evaluation methodology to compare the output of our algorithm against the judgments of human annotators, we demonstrate our results on some relevant examples.

Table 1 shows 6 selected verb argument positions (with argument cases indicated) along with the top ranked HuWN synsets that were identified as preferred semantic classes with our algorithm.

Verbal argument	Semantic class
iszik ACC	*{folyadék}*
to drink sg	{liquid}
kigombol ACC	*{ruha}*
to unbutton sg	{garment}
olvas ACC	*{könyv}*
to read sg	{book}
ül SUP	*{ülőbútor}*
to sit on sg	{seat}
vádol INS	*{bűncselekmény}*
to accuse (sy) with sg	{crime, …}
megold ACC	*{nehézség}*
to overcome sg	{hindrance, …}

Table 1: Automatically identified semantic classes for verb argument positions

We also present the top 5 semantic classes obtained from the nouns found in the accusative argument position of the verb *iszik* (to drink) with their calculated scores in Table 2 (for brevity, we only show the English WN equivalents).

Score	Class	c	d
9.1	{liquid}	26	.35
8.796	{beverage, drink, …}	25	.351
4.888	{alcohol, alcoholic drink, …}	16	.305
4.375	{liquor, spirits, …}	7	.625
3.759	{food, nutrient}	28	.134

Table 2: Top 5 semantic classes identified as direct object arguments of *drink (*c: number of corpus words covered, d: density of the sub-network)

Looking at WN's hierarchy, we see that {liquid} subsumes {beverage, drink} which in turn subsumes {alcohol, alcoholic drink}. But which of these do we exactly want to link {drink} (verb) to? Selecting the most general and most highly ranked category will lead us to choose {liquid}. From a different point of view, however, {beverage, drink} could be more relevant, since not all liquids are drinkable. For some applications indicating the strong association with {alcohol, alcoholic drink} could also be important. By preserving the top *N* semantic classes representing arguments and their degrees of association in the proposed new links, we intend to give an opportunity for future users of our data to freely decide these questions according to their needs.

6 Future Work

Currently we are working on refining our methods. When an evaluation methodology becomes available, it will be possible to fine-tune the candidate scoring formula and to experiment with the best way to assign link probabilities. Additional information that can be used includes corpus frequencies of input words, the depths of the candidate synsets in the hypernym networks and the average distance of the corpus words' synsets from the sub-graphs' root nodes.

As we showed, our method assigns noun frequency lists to verbal argument positions and proposes WN synsets that are most likely to describe selectional preferences. However, argument positions within a verb frame are not independent of each other. It is often the case that binding one of the arguments (assigning a lexical item to that position) entails special selectional preference conditions on another argument position. Examples are *ad ACC* (give something) in the case of ***hírt*** *ad DEL* ("give **news** about sg": to report sg), or *húz ACC* (to pull something) with the argument ***hasznot*** *húz ELA* ("pull **profit** from sg": to profit from sg). As it is also stressed by de Cruys (2010), in the future it is important for us to advance towards a multi-argument model that is able to detect complex verbal units like *hírt ad, hasznot húz* etc. and able to identify selectional preferences for their additional arguments.

According to Mechura (2010), categories in WN do not completely correspond to selectional preferences, and asks the question: "what should an ontology actually look like if it were to reflect accurately the semantic types involved in selectional preferences?" Examining classes that our algorithm assigns with high probabilities may lead to the answer.

7 Conclusion

In this paper, we described a proposed method to automatically enrich Hungarian WordNet with new verb selectional preference relations, which could be useful for semantic text processing tasks. The results may also be beneficial for psycholinguistic research by giving insights to the nature of some of the cross-part-of-speech relationships within the mental lexicon.

Acknowledgments

This work was in part supported by the TÁMOP 4.2.2/B - 10/1-2010-0014 and the TÁMOP 4.2.1.B - 11/2/KMR-2011-0002 projects in the framework of the New Hungarian Development Plan, supported by the European Union, co-financed by the European Social Fund.

References

Brockmann, Carsten -- Lapata, Mirella 2003. Evaluating and combining approaches to selectional preference acquisition. In: Proceedings of EACL 2003, 27-34

Calvo, Hiram -- Gelbukh, Alexander -- Kilgarriff, Adam 2005. Distributional Thesaurus vs. WordNet: A Comparison of Backoff Techniques for Unsupervised PP Attachment. In: Proceedings of CICLing 2005, 177-188

Clark, Stephen -- Weir, David 2002. Class-Based Probability Estimation Using a Semantic Hierarchy. In: Computational Linguistics 28:2, 2002, 187-206

Erk, Katrin 2007. A simple, similarity-based model for selectional preferences. In: Proceedings of ACL 2007, 216-223

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

Guo, Weiwei -- Diab, Mona 2013. Improving Lexical Semantics for Sentential Semantics: Modeling Selectional Preference and Similar Words in a Latent Variable Model. In: Proceedings of NAACL-HLT 2013, 739-745

Kuti, Judit Károly Varasdi Ágnes Gyarmati Péter Vajda 2008. Language Independent and Language Dependent Innovations in the Hungarian WordNet. In Proc. of The Fourth Global WordNet Conference, Szeged, Hungary, 254–268.

Mechura, Michal Boleslav 2010. What WordNet does not know about selectional preferences. In: Dykstra, A. -- Schoonheim T. (eds.) 2010. Proceedings of the 14th Euralex International Congress, Ljouwert/Leeuwarden: Fryske Akademy, 431-436

Miháltz, Márton – Csaba Hatvani – Judit Kuti – György Szarvas – János Csirik – Gábor Prószéky – Tamás Váradi 2008. Methods and Results of the Hungarian WordNet Project. In: Attila Tanács – Dóra Csendes – Veronika Vincze – Christiane Fellbaum – Piek Vossen (szerk.) Proceedings of The Fourth Global WordNet Conference. Szeged: University of Szeged, 311–321.

Prószéky, Gábor 2013. Kutatások egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozás irányában. In: Ladányi Mária ─ Vladár Zsuzsa (ed.) A XI. MANYE-konferencia előadásai (megjelenés alatt)

Prószéky, Gábor – László Tihanyi – Gábor Ugray 2004. Moose: a robust high-performance parser and generator. Proceedings of the 9th Workshop of the European Association for Machine Translation. La Valletta: Foundation for International Studies, 138–142.

Resnik, Philip 1996. Selectional constraints: an information-theoretic model and its computational realization. Cognition 61, 1996, 127-159

Resnik, Philip 1998. WordNet and Class-Based Probabilities. In: Fellbaum (1998a)

Rink, Bryan -- Harabagiu, Sanda 2013. The Impact of Selectional Preference Agreement on Semantic Relational Similarity. In: Proceedings of International Conference on Computational Semantics (IWCS) 2013

Ritter, Alan -- Mausam -- Etzioni, Oren 2010. A latent dirichlet allocation method for selectional preferences. In: Proceedings of ACL 2010, 424-434

Sass, Bálint 2008. The Verb Argument Browser. In: Sojka, P., Horák, A., Kopecek, I., Pala, K. (eds.): 11th International Conference on Text, Speech and Dialog (TSD), Brno, Czech Republic. Lecture Notes in Computer Science 5246, 187-192.

Tian, Zhenhua -- Xiang, Hengheng -- Liu, Ziqi -- Zheng, Qinghua 2013. A Random Walk Approach to Selectional Preferences Based on Preference Ranking and Propagation. In: Proceedings of ACL 2013, 1169-1179

Tufiş, Dan Dan Cristea Sofia Stamou 2004. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In Romanian Journal of Information Science and Technology Special Issue, 7(12), 34.

van de Cruys, Tim 2010. A non-negative tensor factorization model for selectional preference induction. In: Natural Language Engineering 16:4, 2010, 417-437

Váradi, Tamás 2002. The Hungarian National Corpus. In: Zampolli, Antonio (ed.) Proceedings of the Second International Conference on Language Resources and Evaluation. Las Palmas: ELRA, 385-389.

Vossen, Piek 1999. EuroWordNet General Document, Version 3. University of Amsterdam.

Ye, Patrick 2004. Selectional Preference Based Verb Sense Disambiguation Using WordNet. In: Proceedings of the Australasian Language Technology Workshop 2004