First Attempt to Automatically Generate Hungarian Semantic Verb Classes

Bálint Sass[1]

Abstract

Aiming to create verb paraphrases to lay the foundation of sentence paraphrases I automatically created Hungarian semantic verb classes with *k*-means algorithm. The vector representation of verbs was special: dimensions were cases and values were sets of lemmas that can fill the verb frame position defined by the case. I clustered 900 frequent verbs, from which 243 got into 71 smaller clusters, which tend to be semantically coherent. I evaluated the method intuitively, and verified the good classes by contrasting them with a machine readable synonym dictionary, and also by contrasting them with the new Hungarian WordNet.

1. Introduction and related work

We have known it even since Bloomfield – decades before the dawn of computational linguistics – that presumably the most powerful tool in our hands is the distributional analysis. Father of corpus linguistics, John Rupert Firth said that "You shall know a word by the company it keeps." (Firth, 1957) In other words, particular properties of a word can be found if we look at the other words beside it. In her milestone book stated Beth Levin that "the behavior of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large extent determined by its meaning." (Levin, 1993) Accordingly, it can be a fair approach to the meaning of a verb to investigate its complement structure.

Since the above cited hypothesis of Levin – the so called Semantic Base Hypothesis –, there has been a serious effort to investigate the relationship between syntactic behaviour of verbs and meaning of verbs. Verb classes/clusters can be established on the basis of similarities of syntacic behaviour, and these classes can be investigated, whether they are semantically coherent or not. Among the first papers was (Stevenson et al., 1999) or (Schulte im Walde , 2000) , and now there is a comprehensive article available in this field dealing with German semantic verb classes (Schulte im Walde, 2006).

It should be noted that the verb alternations work differently in Hungarian compared to English or German. In general, where English has verb alternations we have different verbs. We can see it even in the active-passive alternation: e.g. *cheer up* is *felvidít* in active and *felvidul* in passive. If we wanted to do what Levin exactly proposed, we should deal with this complicated verb system. Thus, we must forget about the alternations approach, relying only on the hypothesis that similar complement structure entails semantic similarity.

[1] Research Institute for Linguistics, Hungarian Academy of Sciences and
Péter Pázmány Catholic University, Budapest, Hungary
e-mail: joker@nytud.hu

In his paper summarizing principles and mission of corpus linguistics Wolfgang Teubert states that "Meaning is paraphrase." (Teubert, 2005) According to this concept, meaning of a "unit of meaning" is given by the set of its paraphrases. My long-term aim is to collect paraphrases from corpus, and test whether we really can get closer to the meaning of an utterance having all (or some) of its paraphrases. According to Levin there can be a direct feedback from complement structure to the semantic level. Verb classes can automatically be generated based on complement structure similarity. If these classes turn out to be semantically coherent, then we will have similar verbs, in other words verbs, which are paraphrases for each other. These verbs can be the basis of semantic similarity of two sentences: if we have two sentences with two semantically similar verbs and similar complement structures, we can say that these two sentences are paraphrases for each other. Thus, creating semantic verb classes can be a first step towards paraphrase generation.

There is no extensive work available concerning the Hungarian language, but I can not say that my attempt is the first one indeed, because there is another ongoing work somewhat in parallel (Gábor and Héja, 2007). They followed the more traditional approach of applying agglomerative hierarchical clustering to verbs represented by their complement frame distribution. They classified the 150 most freqent verb and the results are very good: they were able to classify 71 verbs (out of 150) into 29 semantically coherent classes according to an intuitive evaluation.

I present another approach, and some preliminary results in this field concerning the Hungarian language.

2. Method and results

Hungarian has about twenty different cases: the case marker at the end of the complement determines the syntactic function of that complement. The fact that the function is morphosyntactically coded, allows broadly free complement order, so order and adjacency of complements play no part. In contrast with SVO languages like English, if a particular Hungarian verb needs a direct object, the accusative-case-marked phrase, which constitutes the direct object can be almost everywhere in the sentence. Thus a Hungarian sentence can be seen as the verb and a *set* of complements. Consequently, position of a complement is defined not by its place in the sentence but morphosyntactically and characterized by the case marker.

To determine verbs and complements of verbs I developed a two step algorithm to process natural language text. The first step was to split up the sentences into clauses: 'clause' here stands for a unit consisting of one verb and its complements. The second step was partial parsing the clauses and determining the complements and two features of them: head-word and case.

I worked out a rule based method for clause detection. The rules are regular expressions, which mark clause borders on the basis of particular punctuation and conjunction patterns. Main principle was that every clause must contain one and only one verb. On the grounds of this principle I added the following supplementary rule: after applying the regular expression rules, if there are two verbs with no clause-border between them, and there are one punctuation mark or one conjunction between them, then the punctuation mark/conjunction will be a clause border (Sass, 2006). The partial parser used for the second step implements cascaded regular grammar technology: grammars are formed from token-level regular expressions, which are built upon each other (Sass, 2005).

Language data comes from the part of speech tagged and disambiguated Hungarian National Corpus (Váradi, 2002), I use the subcorpus of "Magyar Nemzet" daily paper, which consists of eleven million running words.

Based on the above information, I create verb classes using the *k*-means hard clustering algorithm (Schulte im Walde, 2006) in a special way. I take 900 moderately frequent verbs (from the 101st to the 1000th entries from the verb frequency list of the Hungarian National Corpus). An important point is the representation of the verbs. I take the ten most frequent cases and collect all lemmas which occur as head-word of the complement more than five times in these positions with these verbs. Verbs are represented by vectors where the dimensions are the ten most freqent cases and individual values are sets of lemmas that can fill the verb frame position defined by the case. I choose this representation, because this way I do not have to deal with verb frames, just with individual positions. My assumption is that similarity can be found on the basis of lemmas filling syntactic positions beside a verb. With this approach the classical method of complement frame distributions would have leaded to a sparse data problem (Schulte im Walde, 2006).

In the assignment step of the *k*-means algorithm there is a need for a measure of distance between objects to be clustered. Instead of searching the minimum distance between the verb (v) and the means (m) I search for the maximum of 'proximity' defined by the sum of sizes of intersections of the lemma sets in the dimensions:

$$\text{prox}(m,v) = \sum_{c \text{ in case positions}} |m_c \cap v_c|$$

In the update step of the *k*-means algorithm I calculate the new mean according to the following method. For every position I create a frequency list of all lemmas occurring in this particular position for any of the verbs belonging to this mean. I keep only so many lemmas as the average of the lemma count at this position of verbs, and the remainig lemmas will be the most freqent ones.

1. *alkot*, *megalkot* (both: to create)
2. *megtesz*, *megcsinál* (both: to do)
3. *vonatkozik*, *kiterjed* (both: to concern)
4. *meghal* (to die), *megsérül* (to be injured)
5. *függ*, *múlik* (both: to depend)
6. *említ*, *megemlít* (both: to mention)
7. *ismertet* (to outline), *összegez* (to sum up)
8. *módosít* (to modify), *megváltoztat* (to change), *felszámol* (to liquidate)
9. *kiderül* (to turn out), *feltételez* (to assume), *következtet* (to deduce), *bebizonyosodik* (to prove true), *kitűnik* (to get clear)
10. *vizsgál* (to investigate), *tisztáz* (to clarify), *megvizsgál* (to investigate), *elemez* (to analyse), *kutat* (to explore), *feltár* (to reveal)

Table 1: The ten most coherent clusters.

As the *k* value I choose 150, and at the beginning I choose the most 150 verbs as initial means. The convergence was reached after four iterations.

Looking at the resulting clusters I observed that there are some big and several smaller clusters and the smaller ones tend to be semantically more coherent. The ten most coherent clusters can be seen in Table 1.

3. Evaluation

Unfortunately, we do not obtain an exhaustive adequate clustering. It can be seen that there are some plausible classes, and there are types of verbs which usually forms good classes. I evaluate the results in three ways. First I check manually, whether the clusters are intuitively semantically coherent or not. Then I verify the good clusters by contrasting them with a machine readable synonym dictionary, and also by contrasting them with the new Hungarian WordNet.

3.1 Intuitive manual evaluation

As I have mentioned, the smaller clusters tend to be semantically more coherent. I evaluated 71 small clusters with two to six verbs, which cover 243 verbs together. First I checked manually, whether the clusters are semantically coherent or not. This intuitive evaluation results are shown in Table 2. More or less coherent clusters usually contains one "noise" verb, which should not belong to the cluster. It also happens that these clusters should have been separated into two different clusters. These results are to be compared to results of (Schulte im Walde, 2006): about 50 percent coherent, about 25 percent more or less coherent, and about 25 percent not coherent clusters.

coherent clusters	19	27%
more or less coherent clusters	24	34%
not coherent clusters	28	39%

Table 2: Results of the intuitive manual check.

I verify the most coherent ten clusters seen in Table 1 in the following.

3.2 Synonym dictionary based evaluation

I use a machine readable Hungarian synonym dictionary (Kiss, 2001). I check whether the verbs in a cluster occur as synonyms according to the dictionary or not. For clusters with more than two verbs it is to be understood as there is at least one verb pair in the cluster, that occur as synonyms. From the ten good clusters there is only two, which do not occur as a synonym set. This two is the 4. *meghal* (to die), *megsérül* (to be injured) and the 7. *ismertet* (to outline), *összegez* (to sum up) clusters. In the first case there is a graduality, which is an important semantic relation, but which is out of the scope of a synonym dictionary. In the second case, I think that we have a real synonym, which is accidentally missing from the dictionary.

3.3 WordNet based evaluation

Is it not easy to define a good semantic similarity measure based on WordNet, because semantic distance can vary between particular mother-child node pairs in the hypernym-hyponym graph (Patwardhan et al., 2003). One basic notion is the so called *lowest common subsumer*. The lowest common subsumer of two nodes is a node, which is a hypernym of both and none of its hyponym is a hypernym of both. I use the verbal part of the Hungarian WordNet (Kuti et al., 2005). First similarly to the synonym dictionary based evaluation I check whether a cluster occurs as a synset or not. If not I check whether the verbs of a cluster are in hypernym relation, and finally whether they have a lowest common subsumer at least.

Performing this evaluation method I found that from the seven two-verb clusters three can be found as a synset in the WordNet. In other three cases one out of the two verbs can not be found in the WordNet (namely: *megtesz* (to do), *megsérül* (to be injured), and *összegez* (to sum up)) and in the remaining case one verb is in the gloss of the other. In the three bigger clusters there are both same-synset and hypernym relations inside the cluster. It can be said that this evaluation confirms the manual intuitive evaluation.

It should be noted that there is a notion of semantic relatedness which is a broader term compared to semantic similarity (Patwardhan et al., 2003). Semantic relatedness includes e.g. *kind-of*, *part-of*, *opposite-of* relations. It is possible that with the clustering method described above, the semantic relatedness is the thing, which we can capture. There are clusters with opposite meanings (*legyőz* (to defeat), *kikap* (to loose)); with graduality (*meghal* (to die), *megsérül* (to be injured)); or with some specific aspects of an action (*megszűnik* (to cease), *megmarad* (to last), *fennáll* (to exist)).

4. Conclusion and future work

Verbs, which can be called “strong” verbs, have a rich complement structure (i.e. many different lemmas in different case positions) and strong complement structure similarity, tend to get together into a cluster, and such clusters seem to be semantically rather coherent. There are also verbs, which can be called ``weak'' verbs, which occur only with a few cases, and have only a few frequent lemmas in these positions, will get to some big clusters, without any adequate semantic interpretation. It seems that the method described is suitable only for certain verbs, namely the strong verbs mentioned above. It can also be said that there are near-synonyms, which can be captured, as some of the examples in Table 1 show.

Evaluation on the basis of manual intuituve check by native speakers can be good enough, as the two other empirical evaluation method strengthened the adequacy of good classes. The fact that a word is not in a dictionary (or in the WordNet) can not be an argument against the semantic coherence of a verb cluster.

Complement structure can be a good basis of automatic generation of semantically coherent verb classes, although there is a big amount of work to do in this field, and perhaps other clustering methods should be tested. Agglomerative hierarchical clustering can be a better solution as shown by (Gábor and Héja, 2007). Other version of the *k*-means algorithm can be tried, perhaps with splitting up big clusters, and searching for better initialization.

All verb clustering work deals with one-word verbs. An interesting future direction could be to include phrasal verbs, multi-word verbs, to determine e.g. that *megvizsgál* and *górcső alá vesz* (both: to investigate) belong to the same cluster, as do the following two English verbs *to consider* and *to take into consideration*.

References

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. *Studies in linguistic analysis*, pp. 1–32.

Gábor, K. and Héja, E. (2007). Clustering Hungarian verbs on the basis of complementation patterns. *Proceedings of the ACL 2007 conference, Student Research Workshop*, Prague.

Kiss, G. (ed.) (2001). *Magyar Szókincstár*. Tinta Könyvkiadó, Budapest.

Kuti, J., Vajda, P., Varasdi, K. (2005). Javaslat a magyar igei WordNet kialakítására [A proposal for developing the Hungarian WordNet os verbs]. *Proceedings of the 3rd Magyar Számítógépes Nyelvészeti Konferencia [Hungarian Conference on Computational Linguistics] (MSZNY2005)*, Szeged, Hungary, pp. 79–87.

Levin, B. (1993). *English Verb Classes and Alternations.* The University of Chicago Press.

Patwardhan, S., Banerjee, S., Pedersen, T. (2003). Using measures of semantic relatedness for word sense disambiguation. *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico, pp. 241–257.

Sass, B. (2005). Vonzatkeretek a Magyar Nemzeti Szövegtárban [Verb frames in the Hungarian National Corpus]. *Proceedings of the 3rd Magyar Számítógépes Nyelvészeti Konferencia [Hungarian Conference on Computational Linguistics] (MSZNY2005)*, Szeged, Hungary, pp. 257–264.

Sass, B. (2006). Igei vonzatkeretek az MNSZ tagmondataiban [Verb frames in the clauses of the Hungarian National Corpus]. *Proceedings of the 4th Magyar Számítógépes Nyelvészeti Konferencia [Hungarian Conference on Computational Linguistics] (MSZNY2006)*, Szeged, Hungary, pp. 15–21.

Schulte im Walde, S. (2000). Clustering verbs semantically according to their alternation behaviour. *Proceedings of the 18th International Conference on Computational Linguistics (COLING2000)*, Saarbrücken, Germany, pp. 747–753.

Schulte im Walde, S. (2006). Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2), pp. 159–194.

Stevenson, S., Merlo, P., Karieva, N., Whitehouse, K. (1999). Supervised learning of lexical semantic verb classes using frequency distributions. *Proceeding of SIGLEX'99*, College Park, Maryland.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10(1), pp. 1–13.

Váradi, T. (2002). The Hungarian National Corpus. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, Las Palmas, Spain, pp. 385–389.