Detecting optional arguments of verbs

András Kornai, Dávid Márk Nemeskey
HAS Computer and Automation Research Institute
H-1111 Kende u 13-17, Budapest
{kornai,nemeskey}@sztaki.hu

Gábor Recski
HAS Research Institute for Linguistics
H-1068 Benczur u 33, Budapest
recski@mokk.bme.hu

Abstract

We propose a novel method for detecting optional arguments of Hungarian verbs using only positive data. We introduce a custom variant of collexeme analysis that explicitly models the noise in verb frames. Our method is, for the most part, unsupervised: we use the spectral clustering algorithm described in Brew and Schulte in Walde (2002) to build a noise model from a short, manually verified seed list of verbs. We experimented with both raw count- and context-based clusterings and found their performance almost identical. The code for our algorithm and the frame list are freely available at http://hlt.bme.hu/en/resources/tade.

Keywords: optional arguments, spectral clustering, collexeme analysis

1. Introduction

One of the classical puzzles in linguistics is to make the distinction between obligatory verbal arguments as in *John likes broccoli* which cannot be omitted (**John likes*) and optional arguments as in *John eats broccoli* where removal of the argument results in a sentence that is less informative (*John eats*) but still grammatical. There is nothing in the semantics that would make this distinction obvious: there can be no act of *liking* that doesn't involve *liking something* just as there can be no act of eating that doesn't involve *eating something*. Note that the lack of bare *like* is negative evidence, not directly present in primary linguistic data. Even in large corpora, the empirical frequency of very rare but grammatical constructions is zero, making them indistinguishable from constructions that are ungrammatical, an observation that has led many linguists (starting with Chomsky 1957) to the conclusion that one must rely on introspection to tell the two apart.

Stefanowitsch (2006) uses 2x2 contingency tables and the standard Fisher exact test to show that negative evidence can be meaningfully extracted from large labeled corpora especially for features that are "relatively uncontroversially tagged". Unfortunately, the recall of this method is greatly limited by the size of the gold data: for example Li and Abe (1999) or Gábor and Héja (2007) relied on manually tagged gold corpora (the Penn and Szeged Treebanks respectively) to obtain results for 354 English (resp. 150 Hungarian) verbs. Here we extend the less supervised approach pioneered by Korhonen (1998) for English and Sass (2010) for Hungarian: we take large and unanalyzed corpora, extract patterns by shallow parsing (Briscoe 1997), and postprocess the results. The standard approach is to set rather high thresholds, in effect trading in recall for acceptable precision: for example Sass (2010) keeps only patterns that occur over 250 times in the data, yielding 2,200 verbs and 175 frames. Strong thresholding, however, destroys the sensitivity to optional arguments, since lack of above-threshold evidence for the intransitive pattern is not evidence for lack of this pattern.

In terms of learning an embedding, the unsupervised method would amount to running standard algorithms such as word2vec or GloVe on the output of the preprocessor, with the context of a verb defined as case-marked NPs, PPs, *that*-clauses, and infinitival clauses occurring in a clause. In order to avoid many of the issues that arise in the setting of hyperparameters (Levy et al. 2015) we concentrate on the algorithmic core, the spectral clustering of the data. Spectral methods were pioneered for verb frame clustering by Brew and Schulte im Walde (2002) in a supervised context, with the frames induced in an earlier pass (Schulte im Walde 2002) using PCFGs.

Here we work on the unsupervised task, with induction of the frames and their clustering performed in the same pass. The main novelty is a more sensitive thresholding technique, which improves the yield notably (we derive 377 high quality frames for 3,300 verbs), making the resulting set robust enough for obtaining negative conclusions as well. The rest of this introduction describes the main data sources. Section 2. describes and evaluates the results of the spectral clustering against other clustering methods and against a manually encoded gold standard dataset. Automatic acquisition of the obligatory/optional distinction is discussed in Section 3..

Hungarian nouns may be marked for one of 19 cases (many of these would be marked by prepositions in English). In addition to cases, we also considered 161 types of postpositional phrases (PPs) and the subordinating conjunction *hogy* 'that', which indicates a clausal argument, for a total of 181 complement types. Finally, if an infinite verb is present in a clause, it is treated as another complement type (an argument of the finite verb), while all NPs and PPs are considered arguments of the infinitive: for example in *John wants to drink beer* we would take *to drink* as the argument of *want*, and *beer* as the argument of *drink*.

Our main dataset is based on the Hungarian Webcorpus (Halácsy et al., 2004), which contains over 42 million sentences gathered from the .hu domain. Morphological analysis and the identification of maximal syntactic phrases were performed using the hunmorph (Trón et al., 2005) and hunchunk (Recski and Varga, 2009) tools respectively. We extracted frame patterns from all clauses in the corpus which contain exactly one finite verb, assigning to its case frame all top-level NPs and PPs. Altogether we an-

alyzed 58.9 million clauses. Given the free phrase order of Hungarian, a unique pattern is defined as a verb followed by the sorted list of arguments – there are over 136k distinct patterns in the data. In order to reduce noise caused by errors in morphological analysis and chunking, we applied two levels of filtering. In the `web_50` set patterns containing verbs with fewer than 50 occurrences were discarded similar to Li and Abe (1999), and in the `web_250` set we discarded all patterns with absolute frequency below 250, similar to Sass (2010, 2011).

A reference dataset was kindly provided by Bálint Sass, who used rule-based methods to extract verb frames from the Hungarian National Corpus (Váradi 2002). The `Sass` dataset was also cut off at 250. Finally, we considered the manually created case frames presented in Papp (1969), which covers all verbs listed in the seven-volume Explanatory Dictionary of the Hungarian Language (Országh 1962). Neither of these sources include PPs or infinitival complements, and Papp further collapses some cases such as the inessive *ban* and the superessive *on*, or the illative *ba* and the superessive *ra*, in single codes. The main parameters of the four sets (number of clauses, verb–frame patterns, verb and frame types) are summarized in Table 1 below.

dataset	clauses	patterns	verbs	frames
web_50	58.9M	1.1M	13.7k	136.8k
web_250	45.8M	20k	4.3k	944
Sass	6.1M	6.2k	2.2k	175
Papp	n/a	n/a	15k	128

Table 1: Summary of case frame datasets

Both our frame list and the Papp frames, as well as the code of our algorithm are freely available at `http://hlt.bme.hu/en/resources/tade`.

2. Creating and evaluating the clusterings

Both our clustering techniques and our evaluation measures are similar to those employed by Brew and Schulte im Walde (2002), except we took advantage of the lessons they learned and omitted binary cosine similarity from the list of divergences considered for clustering, trying Euclidean (L2), Kullback-Leibler (KL), and Jensen-Shannon (JS) divergences in addition to the skew and cosine measures they worked with. We performed direct clustering only as a sanity check, and confirmed their result that the Ng et al. (2001) spectral clustering algorithm obtains better clusters. In the *long vector* (lv) condition the clusterings were done on sparse vectors of approximately 136k dimension that listed, for each frame, the absolute frequency of the verb in that frame. In the *short vector* (sv) condition the clustering was based on vectors of 182 dimension that simply listed the absolute frequency of each argument type with each verb, irrespective of frames.

Before turning to optional arguments, we needed to assess the quality of the clustering output. As we shall see, we induced frames that are reasonable in terms of precision, and vastly improve recall, so that in terms of F-measure they represent considerable progress over earlier work on Hungarian. We first computed *alignment* (first suggested as a measure of clustering similarity by Cristianini et al. 2001), between the manually coded Papp dataset and the output of the automatic clusterings. We varied several parameters, including the distance measure, the σ value of the clustering algorithm, and the number of clusters k. Table 2 summarizes the parameters of the best alignments for $k = 128$ (ignore the last column for now).

Dataset	σ	dist	cond	align	VI
web_50	.1	L2	sv	0.500	6.361
web_50	.1	L2	lv	0.498	6.621
web_50	.1	cos	lv	0.497	6.644
web_50	1.0	cos	sv	0.496	6.707
web_250	1.0	cos	lv	0.541	5.425
web_250	.1	cos	lv	0.540	5.416
web_250	.1	cos	sv	0.540	5.475
web_250	.01	cos	sv	0.539	5.516
Sass	.1	cos	lv	0.533	4.360
Sass	.01	cos	sv	0.531	4.470
Sass	1.0	cos	lv	0.531	4.368
Sass	.01	cos	lv	0.531	4.398

Table 2: Alignment with Papp data

These numbers are not nearly as good as those reported by Brew and Schulte im Walde (2002) for alignment between their clustering and their gold data, which were in the 0.80-0.86 range. However, the number of clusters used in this test is an order of magnitude larger (128 vs. 14), and our gold set was created by a far less sophisticated methodology than theirs (Papp simply asked his students to code the frames directly).

Perhaps the most surprising result evident from this table is that the short vector and the long vector conditions are quite competitive: this lends support to a recent finding of (Levy et al. 2015, Stratos et al. 2015) that 'count-based' methods are not necessarily inferior to 'predictive' methods. Not as surprising, but quite robust is the observation that cosine distance works best, with L2 becoming useful only on the largest and most noisy `web_50` dataset. In what follows we restrict ourselves to cosine similarity. Since this already has the right properties to serve as a direct measure of affinity (all coordinates are non-negative, and so are the scalar products of both long and short vectors), the step of computing the affinity matrix A from the distance matrix can be omitted entirely. Besides making the computation simpler, using the cosine similarity directly for A has the additional benefit that we no longer need to search the space for the best value of σ.

We checked the robustness of our clusterings against each other as well. Rerunning clusterings with different random seeds but no change of parameters produces clusterings that align 0.98 or better, and changes in σ have similarly negligible effect. The best alignments across different datasets are in the 0.85-0.92 range: for example `Sass` long vectors against `web_250` short vectors give 0.923, against `web_50` long vectors give 0.922, and `web_50` against `web_250` is 0.851. When larger numbers of clusters are considered, the numbers become even better: for example 1024 clusters based on the `Sass` data with short vs. long vectors

align within 0.993, Sass against web_250 (short vectors) also within 0.993.

Since the automatically generated clusterings are considerably closer to one another than to the Papp data, we need a measure more sensitive than alignment to investigate the reasons for the discrepancy. In principle, Papp's coding system could distinguish 1,690 verb frames, but the 14,988 verbs considered by him populate only 128 of these. The spread across these frames is very uneven, the entropy of the distribution is only 2.64 bits based on type frequencies, token frequency weighted entropy is 2.78. The reason is that the gold data is dominated by 4,743 intransitive and 6,257 transitive verbs, together accounting for over 73% of types (71% of tokens weighted by Webcorpus frequency), while the distribution of clusterings is far more even, with entropies over 6.5 (the theoretical maximum for a perfectly even clustering would be 7 bits for 128 clusters). It therefore makes sense to also compare clusterings based on the *variation of information* measure of $VI(\mathcal{C}, \mathcal{C}')$ (Meilă, 2003), which is more robust to cluster size and number of clusters (see Christodoulopoulos et al. 2010) – these are the numbers in the last column of Table 2.

3. Learning optionality

Until now, everything we did was unsupervised, but for detecting optionality we need a bit of weak supervision as follows. The large intransitive and transitive categories in the gold data are divided in several clusters by the automatic method: what we need is to find the 'true inheritor' of these classes among the automatically created clusters. For example, when we trace the 945 gold intransitives that appear in the web_50 clustering (here and in what follows numbers are taken from one specific run, since the variance across runs is negligible), 549 of these land in a single cluster containing verbs like *tüsszög* 'sneeze' and *ácsorog* 'stand in one place, loiter'. Similarly, tracing the 1,661 gold transitives leads to a major cluster of 514 verbs such as *hatástalanít* 'disarm' and *kisajátít* 'monopolize'.

We emphasize that finding these two clusters is essentially an automatic process requiring only minimal human supervision, feasible even in languages where no gold data similar to Papp (1969) is available: we just need to spot-check a few dozen words in the largest automatically induced clusters to find the main intransitive and transitive clusters, since these contain for the most part words that one would unambiguously classify as (in)transitive. We will rely on these two clusters to provide us with background statistics to help us set apart true case frames and optionality from noise in the data.

Let us summarize what we have so far. First, we have obtained robust clusterings for $k = 128, 256, 512, 1024$ based on differently prepared datasets (Sass, web_50, web_250) using both short and long feature vectors. Second, we have a large number (over a hundred thousand) potential patterns such as *változtat*+ACC+során 'change sg during sg'. Intuitively it is clear that the PP[during] element is a free adverbial, only the accusative NP, the object of *change*, is part of the case frame. Third, we need to decide whether the argument is really obligatory, distinguishing cases like *eat,* where the ACC is part of the frame but can be omitted, from cases where the optionality stems from not being part of the frame to begin with, as with the adverbial of circumstance above. The data, needless to say, is very noisy, and we need a good model of this noise: this is what the automatically obtained intransitive I and transitive T clusters provide.

Given a verb V such as *változtat* and a putative frame F such as ACC+során we begin with four numbers arranged in a 2 by 2 table:

	F	$\overline{F}$
V	53	33,325
I	8	403,173

We have 53 occurrences of the verb in the frame, and 33,325 outside the frame – this much is standard. The novel element, compared to regular collexeme analysis (Stefanowitsch, 2003), is that we contrast these to the 8 occurrences of I-verbs in the frame to the 403,173 occurrences of intransitive verbs outside the frame, rather than to a baseline of all verbs. The point is that we accept the verbs in class I as true intransitives, and treat their every occurrence in some frame (other than the empty frame) as pure noise.

Since we cannot guarantee that all elements are above 5, we use Fisher's exact test to determine whether the ratio 53/33325 is significantly above the baseline 8/403173, using a p-value of 10^{-75}. We use Stirling's approximation to compute the value. The threshold may appear unusually strict, but works quite well in practice. In our example, we get $p \sim 10^{-50}$ and the frame is rejected.

With this test, we accept 377 frames (3,297 verbs) for a total of 21,718 patterns. We declare an element Y of an accepted frame X *optional* if $X \setminus Y$ is also an accepted frame. In particular, if the empty frame is accepted (this may happen to some words outside the initial I cluster), any single argument will be by definition optional. To give an example, the verb *megtalál* 'find' has 32 accepted frames before this reduction step, but only 21 afterwards, since a complex frame such as ACC+(ALL) now stands for two frames ACC and ACC+ALL. On average, a quarter of the reduced frames contains an optional element.

Since they take up over 70% of the probability mass, the intransitive and transitive categories deserve special attention. Intransitives are defined as those verbs that (i) have the empty frame among their significant frames and (ii) have no other significant frame. This is not exactly the same as being a member of I, the largest cluster with typical intransitives, but the precision (recall) of the two sets relative to the gold data is about the same, 58% (62%) for an F-measure of 0.60. Transitives are defined by (i) having the ACC frame and (ii) having no other frame, except perhaps the empty frame, in which case we say the object is optional, as in the verb *eat*. The cluster T has higher precision (75%) but lower recall (31%) for a combined $F = 0.44$. These numbers compare rather favorably to the F-measures obtained by evaluating the Sass (2011) results against the same gold standard, $F = 0.05$ (intransitives) and $F = 0.12$ (transitives).

Turning to the rest of the data, the web_50 set has data for 3,297 of those 3,978 verbs that Papp considers neither

intransitive nor transitive. Compared to the gold frames, our precision is 22%, recall 29%, for $F = 0.25$ – Sass obtains $F = 0.095$ on these. To appreciate these numbers, it should be noted that earlier work was restricted to a few hundred examples, while the results of Sass (2011) and our work are measured here against the entire headword list of the 7-volume Explanatory Dictionary, nearly 15k verbs.

4. Conclusions

We have presented a high yield high precision algorithm for the extraction of case frames with optional elements. It is only because we actively model the noise in the computation that we can use lower thresholds (50 verb occurrences compared to 250 pattern occurrences used in earlier work), improving recall to the point that detection of optionality becomes possible.

We see this as a step toward the eventual goal of extracting deep case relations from the data. The task is twofold: first, to extract a clustering that has the detail of the Levin classification, about a thousand clusters, and second, to *organize* these clusters in a meaningful way. The kind of surface case frames presented here would need to undergo clustering themselves to recover the wisdom of Papp (1969) that the illative *ba* and the sublative *ra* can be actually coded together as they both express a deeper GOAL relationship between the main verb and the argument. Whether the terminative *ig* is also part of this cluster is a question that we should be able to settle empirically. Given that the information variation across our $k = 1024$ clusterings is less than 3 bits, we may be within striking distance of this goal.

Acknowledgments

We thank Bálint Sass (HAS RIL) for the `Sass` dataset, and Sascha Griffiths (Queen Mary University of London) and the LREC referees for their comments.

5. Bibliographical References

Brew, C. and Schulte im Walde, S. (2002). Spectral clustering for German verbs. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 117–124. Association for Computational Linguistics.

Briscoe, E. J. and Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ANLP Conference*, pages 356–363.

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague.

Christodoulopoulos, C., Goldwater, S., and Steedman, M. (2010). Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 575–584. Association for Computational Linguistics.

Cristianini, N., Shawe-Taylor, J., and Kandola, J. (2001). Spectral kernel methods for clustering. *Advances in neural information processing systems*, 14:649–655.

Gábor, K. and Héja, E. (2007). Clustering Hungarian verbs on the basis of complementation patterns. In *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*, pages 91–96. Association for Computational Linguistics.

Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., and Trón, V. (2004). Creating open language resources for Hungarian. In *Proc. LREC2004*, pages 203–210.

Korhonen, A. (1998). Automatic extraction of subcategorization frames from corpora - improving filtering with diathesis alternations. In *Proceedings of the ESSLLI 98 Workshop on Automated Acquisition of Syntax and Parsing*, pages 49–56.

Levy, O., Remus, S., Biemann, C., and Dagan, I. (2015). Do supervised distributional methods really learn lexical inference relations? In *NAACL*.

Li, H. and Abe, N. (1999). Learning dependencies between case frame slots. *Computational Linguistics*, 25:283–291.

Meilă, M. (2003). Comparing clusterings by the variation of information. In Bernhard Schölkopf et al., editors, *Learning theory and kernel machines*, pages 173–187. Springer.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856. MIT Press.

Országh, L. (1959-62). *Explanatory Dictionary of the Hungarian Language*. Akadémiai, Budapest.

Ferenc Papp, editor. (1969). *A magyar nyelv szóvégmutató szótára*. Akadémiai Kiadó, Budapest.

Recski, G. and Varga, D. (2009). A Hungarian NP Chunker. *The Odd Yearbook. ELTE SEAS Undergraduate Papers in Linguistics*, pages 87–93.

Sass, B., Váradi, T., Pajzs, J., and Kiss, M. (2010). *Magyar igei szerkezetek. A leggyakoribb vonzatok és szókapcsolatok szótára.* Tinta.

Sass, B. (2011). *Igei szerkezetek gyakorisági szótára*. Phd thesis, Péter Pázmány Catholic University.

Schulte im Walde, S. (2002). A subcategorisation lexicon for German verbs induced from a lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*.

Stefanowitsch, A. and Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2):209–243.

Stefanowitsch, A. (2006). Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory*, 2(1):61–77.

Stratos, K., Collins, M., and Hsu, D. (2015). Model-based word embeddings from decompositions of count matrices. In *Proceedings of 53rd ACL*.

Trón, V., Gyepesi, G., Halácsy, P., Kornai, A., Németh, L., and Varga, D. (2005). Hunmorph: open source word analysis. In Martin Jansche, editor, *Proceedings of the ACL 2005 Software Workshop*, pages 77–85. ACL, Ann Arbor.

Váradi, T. (2002). The Hungarian national corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 385–389.