A Unified Method for Extracting Simple and Multiword Verbs with Valence Information and Application for Hungarian

Bálint Sass
Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
sass.balint@nytud.hu

Task

sentence	construction	structure
(1) "I believe in miracles."	x believes in y	free in-slot
(2) "The girl shrugs her shoulder."	z shrugs (z's) shoulder	fixed object slot
(3) "Nobody takes that into consideration."	u takes v into consideration	free object slot + fixed into-slot

Free slots correspond to *valences*. Fixed slots plus the verb form a *multiword verb*.

Some constructions (e.g. take something into consideration) show both properties – being multiword and having valence at the same time.

Definition: *verb-centered constructions (VCCs)* are simple and multiword verbs, with or without valence: verb + zero or more valences + zero or more additional NPs/PPs with fixed content words. The verb together with the NPs/PPs (if any) must have non-compositional/idiomatic meaning.

Aim: to develop a method which extracts all kinds of VCCs from corpus, those also which are multiword and have valence at the same time.

Motivation. Although full-grown VCCs are frequent, they usually get out of field of vision being a borderline case between MWEs and verb subcategorization frames (SCFs). We need lexical acquisition methods to handle them also.

Method

Corpus representation. Basic unit: clause = verb + its dependents
Dependent representation: slot name + ':' + the lemma of the head (if fixed)

Repr. of sentence (3), its *clause skeleton (CS)*:	take SUBJ:nobody OBJ:that into:consideration	← *INPUT*
Repr. of its VCC:	take SUBJ OBJ into:consideration	← *OUTPUT*

Processing the corpus clauses the algorithm should detect:
(1) which slot is integral part of the VCC, (2) whether a dependent slot is free or fixed.

Main idea: we store initially all slots and all content words and allow the algorithm to get rid of
(1) complete slots, when they are not integral part of a VCC;
(2) content words, where they are just filling in a valence slot.

1. We take all CSs of the corpus with frequency counts.
   We perform *alternating omission* that means we add some "free" CS variants to the initial list:

VCC	length
shrug SUBJ:girl OBJ:shoulder →	4
shrug SUBJ OBJ:shoulder	3
shrug SUBJ:girl OBJ	3
shrug SUBJ OBJ	2

   This step makes possible to have VCCs with free slots in the resulting list.
2. We sort the resulting verb frame list according to *length* which is: number of free slots + number of fixed slots · 2.
3. Starting with the longest one we discard CSs with frequency less than 5, and add their frequency to a *one-unit-shorter* frame on the list. If there are several such frames, we choose randomly.
4. Intended VCCs are the final remaining verb frames, ranked by cumulative frequency.

"take SUBJ OBJ into:consideration" will be on the resulting list because in the corpus clauses whose main verb is take,
- the into slot is usually filled by the word consideration – so its frequency can cumulate,
- but the OBJ slot is much more variable – so words in this slot are more easily dropped out.

The method is a generalization of a former subcategorization frame extraction method. *Our contribution:* the idea of storing all content words, the alternating omission procedure, and the suitable definition of length for VCCs.

The representation and the method is in essence *language independent*, it could be applied to other languages as well.

Evaluation

Corpus: Hungarian National Corpus.
Slots: Hungarian casemarks.
Required representation is obtained from an automatic dependency annotation of the HNC.
Evaluation method:
n-best lists with two annotators.

Annotation criterion. According to the definition of VCCs the annotation criterion was this: a candidate is a true positive VCC if and only if (1) there is no fixed slots or the verbal part (verb + occurrent fixed slots) has non-compositional/idiomatic meaning; and (2) the (possibly multiword) verb truly has such a subcategorization frame which is present and this frame is complete.

Results. Average precision values by type and by n (of n-best list). The $\pm$ percentages point out two values corresponding to the two annotators. Cohen's κ corresponds to the rightmost percentage value in every row. In the 'total' line we evaluate the first 500 candidates of the whole list.

type	$n=$ 50	100	150	200	500	Cohen's κ
no slots	83.0% ± 5.0%	82.0% ± 4.0%				0.53
1 free slot	94.0% ± 2.0%	92.0% ± 1.0%	92.0% ± 0.7%	91.8% ± 0.8%		0.77
↪ object	99.0% ± 1.0%	97.0% ± 1.0%	98.0% ± 0.7%	98.0% ± 0.5%		0.75
↪ other	79.0% ± 1.0%	79.5% ± 0.5%	78.7% ± 1.3%	79.8% ± 1.8%		0.68
1 fixed slot	58.0% ± 6.0%	44.0% ± 3.0%				0.64
↪ subject	20.0% ± 6.0%	19.0% ± 6.0%				0.43
↪ other	83.0% ± 1.0%	80.5% ± 1.5%				0.33
2 free slots	77.0% ± 7.0%	66.5% ± 8.5%				0.63
1 free + 1 fixed	94.0% ± 0.0%	88.5% ± 3.5%	87.0% ± 3.0%	83.3% ± 3.3%		0.59
2 fixed slots	51.0% ± 7.0%	39.0% ± 5.0%				0.50
total	94.0% ± 0.0%	93.5% ± 1.5%	89.3% ± 1.3%	89.5% ± 1.5%	88.9% ± 1.3%	0.65

Discussion.
– 1 free slot: highest inter-annotator agreement. Precision values coming close to 100 percent in the case of simple transitive verbs.
– 1 non-subject fixed slot: precision values above 80 percent, but low κ values.
– *Main result.* 1 free + 1 fixed slot (namely multiword verbs with one valence slot): precision values above 80 percent, moderate inter-annotator agreement. Significance of our approach lies in its capability to deal with multiword verbs and their valence simultaneously, handling SCFs and MWVs in a uniform general way.

Examples

First ten real VCCs with 1 free + 1 fixed slot from the resulting list. 'X' means that the particular Hungarian casemark do not have an exact English counterpart.

1. van szó -rÓl
   be word-SUBJ ABOUT
   ≈ 'something is said'
2. tesz lehető-vÁ -t
   make possible-X OBJ
   'make something possible'
3. van szükség -rA
   be need-SUBJ ONTO
   'something is wanted'
4. vesz ész-rA -t
   take mind-ONTO OBJ
   'became aware of something'
5. kerül sor -rA
   come line-SUBJ ONTO
   ≈ 'something takes place'
6. vesz figyelem-bA -t
   take consideration-INTO OBJ
   'take something into consideration'
7. hoz lét-rA -t
   bring being-INTO OBJ
   'create something'
8. tart fontos-nAk -t
   hold important-FOR OBJ
   'think something is important'
9. vesz rész-t -bAn
   take part-OBJ IN
   'take part in something'
10. vesz tudomás-Ul -t
    take knowledge-X OBJ
    'take notice of something'

RANLP 2009 Conference, 14-16 September 2009, Borovets, Bulgaria