First attempt to automatically generate Hungarian semantic verb classes

Bálint Sass
sass.balint@itk.ppke.hu

Péter Pázmány Catholic University, Budapest, Hungary

4th Corpus Linguistics Conference
27-30 July 2007, Birmingham

1. Introduction
2. Method
3. Results
4. Evaluation
5. Conclusions

1. Introduction
2. Method
3. Results
4. Evaluation
5. Conclusions

Introduction

“You shall know a word by the company it keeps.”

(John Rupert Firth)

“. . . 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.”

(Beth Levin)

VERB ALTERNATIONS

Are there any verb alternations in Hungarian?

English: active passive alternation – Hungarian: different verbs

EXAMPLE

cheer up = *felvidít* (in active), *felvidul* (in passive)

→

Hypothesis:
similar complement structure entails semantic similarity.

PARAPHRASE AND MEANING

“Meaning is paraphrase.”

(Wolfgang Teubert)

Aim:

1. collect paraphrases from corpus
2. test whether we get closer to meaning having all (or some) paraphrases

Paraphrase and meaning

Semantic Base Hypothesis:
complement structure → semantic level

A method for identifying paraphrases:

1. complement structure similarity → automatically generated verb classes
2. semantically coherent classes? → verb-paraphrases
3. two sentences with two semantically similar verbs and similar complement structures → paraphrases

NOT FIRST . . .

No extensive work in this field for Hungarian.

Kata Gábor and Enikő Héja:
Clustering Hungarian Verbs on the Basis of Complementation Patterns (ACL 2007, Student Research Workshop)

- *verb representation:* complement frame distribution vector
- *algorithm:* agglomerative hierarchical clustering
- 150 most frequent verbs
- *results:* 71 verbs in 29 semantically coherent classes according to an intuitive evaluation

HUNGARIAN VERBS AND COMPLEMENTS

- Hungarian: twenty different cases
- case marker – determines syntactic function
- → free complement order
- simple Hungarian sentence: verb + a *set* of complements
- morphosyntactic complement *positions*

Determining verbs and complements

Two step algorithm:

1. sentences → clauses
   clause = verb + its complements

   — regular expression rules

2. partial parsing → complements: head-word and case

   — cascaded regular grammar for NPs

1. Introduction
2. Method
3. Results
4. Evaluation
5. Conclusions

LANGUAGE DATA

- 11 million running words
- “Magyar Nemzet” daily paper
- part of the Hungarian National Corpus
- POS-tagged & disambiguated

Representation of verbs

k–means clustering algorithm

- verb – vector
- *dimensions:* ten most frequent cases
- *values:* sets of lemmas

Example representation

	vonatkozik (to concern)
NOM	*szabály* (rule), *törvény* (law)
ACC	–
DAT	–
INE	–
SUB	*ők* (they), *mindenki* (everybody), *épület* (building)

Representation of verbs

k–means clustering algorithm

- verb – vector
- *dimensions:* ten most frequent cases
- *values:* sets of lemmas

Example representation

	összegez (to sum up)
NOM	*elnök* (president)
ACC	*tapasztalat* (experience), *eredmény* (result)
DAT	–
INE	–
SUB	–

k-MEANS: ASSIGNMENT STEP

- need for distance measure between verbs
- *proximity:* sum of sizes of intersections of the lemma sets

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

m – mean, v – verb

k-MEANS: UPDATE STEP

To calculate the new mean …

- for every dimension:
  frequency list of all lemmas for all of the verbs belonging to this mean
- keep the most frequent lemmas
- keep so many lemmas as the average of the lemma count at this position of verbs

k-MEANS

- 900 moderately frequent verbs
- *k* (number of clusters) = 150
- *initialization:* most frequent 150 verbs
- *convergence:* reached after four iterations

RESULTS

- 51 single-verb clusters
  71 smaller (2 to 6 verbs) clusters: 243 verbs
  28 bigger clusters
- smaller clusters are semantically more coherent
  algorithm was able to cluster these verbs
- evaluation – only the smaller clusters

Results

The ten most coherent clusters:

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)

EVALUATION METHODS

Three ways:

1. manual intuitive check
2. verify most coherent clusters: synonym dictionary
3. verify most coherent clusters: Hungarian WordNet

Manual evaluation

Results of the intuitive manual check:

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

Common errors:

- coherent cluster with one “noise” verb
- two separate coherent clusters mixed up

Verification – synonym dictionary

- a machine readable Hungarian synonym dictionary: “Magyar Szókincstár”
- Are verbs in a cluster synonyms?
  yes: 8 ↔ no: 2
- Clusters not verified:
  meghal (to die), *megsérül* (to be injured)
  ismertet (to outline), *összegez* (to sum up)

VERIFICATION – HUNGARIAN WORDNET

- verbal part of the new Hungarian WordNet
- Do verbs in a cluster appear in the same synset?
  If not, are they at least in hypernym relation?
  - 7 two-verbs clusters:
    3 found as a synset
    3 – missing verb
    1 verb is in the gloss of the other
  - 3 bigger clusters:
    both same-synset and hypernym relations

1. Introduction
2. Method
3. Results
4. Evaluation
5. Conclusions

fides et ratio

Conclusions

- two empirical evaluation methods strengthened the results of the manual intuitive evaluation
- no argument can be based on missing words
- capable of capturing similar verbs with *rich* complement structure
- capable of capturing near-synonyms

Conclusions

semantic relatedness: kind-of, part-of, opposite-of . . .

Example – opposite meaning

legyőz (to defeat), *kikap* (to loose)

Example – graduality

meghal (to die), *megsérül* (to be injured)

Example – specific aspects of an action

fennáll (to exist), *megszűnik* (to cease), *megmarad* (to last)

Future work

- agglomerative hierarchical clustering can be a better solution

- other versions of the algorithm
  – splitting up big clusters
  – better initialization

- include phrasal verbs, multi-word verbs

Example

megvizsgál, *górcső alá vesz* (both: to investigate)

Example

to consider, to take into consideration

Thank you for your attention!