Wrocław University of Technology

Corpus-based Semantic Relatedness for the Construction of Polish WordNet

Bartosz Broda[1], *Magdalena Derwojedowa*[3], *Maciej Piasecki*[1], *Stanisław Szpakowicz*[2],

1. Institute of Applied Informatics, WUT
2. Institute of the Polish Language,Warsaw University
3. School of Information Technology and Engineering, University of Ottawa

plwordnet.pwr.wroc.pl

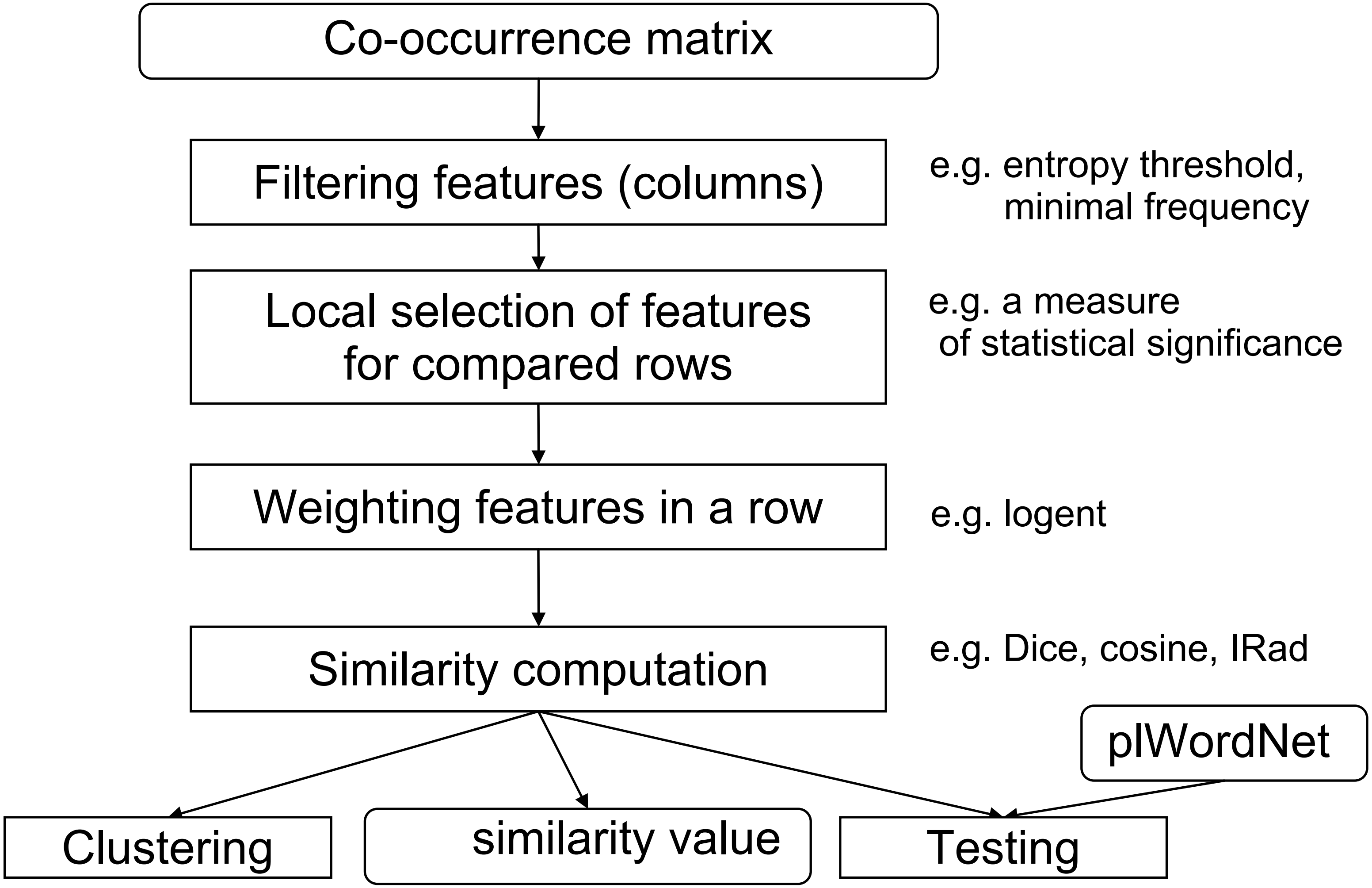
Plan

- Measure of Semantic Relatedness (MSR) in Building a Wordnet
- Rank Weight Function as the Basis for MSR
- Lexico-morphosyntactic Constraints
- Experiments and WordNet-Based Synonymy Test
- MSR and Wordnet Extensions
- Observations and future work

MSR in Building a Wordnet

- High linguistic workload makes wordnet construction very costly
  - assumption: automatic acquisition of lexico-semantic relations can reduce the cost
- MSR: $LU \times LU \to R$
    - pairs of lexical units are mapped into real numbers
    - a lexical unit – a lexeme or a multiword expression
  - LUs semantically related to some LU should receive significantly higher values than unrelated LUs

Framework for MSR



Co-occurrence matrix

Filtering features (columns)

e.g. entropy threshold,
minimal frequency

Local selection of features for compared rows

e.g. a measure of statistical significance

Weighting features in a row

e.g. logent

Similarity computation

e.g. Dice, cosine, IRad

plWordNet

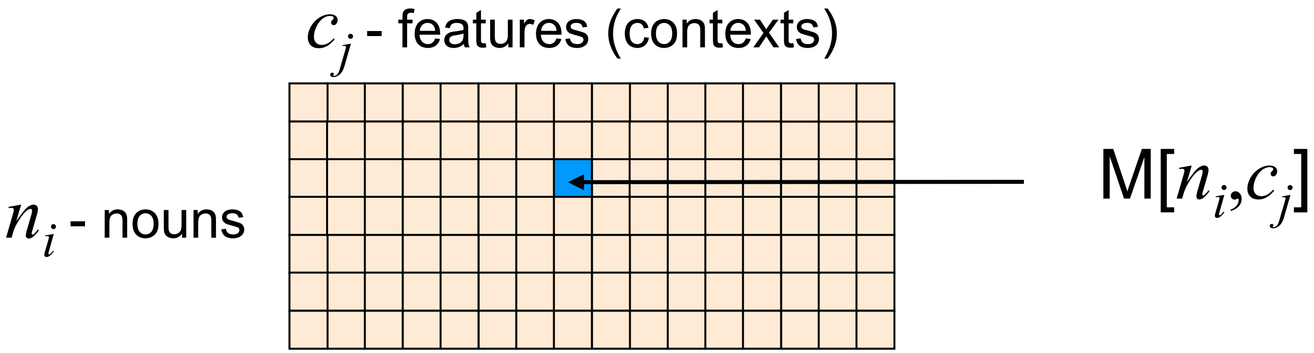
Clustering

similarity value

Testing

Co-occurrence Matrices

- Scheme



c_j - features (contexts)

n_i - nouns

M[n_i,c_j]

- Typical characteristics:
  - very large size: many thousands × many thousands
  - sparsity
  - substantial level of noise, e.g. accidental frequencies
- Features:
  - documents or paragraphs
  - co-occurrence in a text window

Rank Weight Function

- Problem with normalising values of MSR
  - feature values depend on frequency
  - no corpus is perfectly balanced
  - different weighting function did not solve the problem
- The need for generalisation from frequencies
  - not all the features are significant discriminators for every pair of nouns
  - ranking of relative importance of features instead of raw counts

Rank Weight Function

- Algorithm of transformation
  1. Weighted values of the cells are recalculated using a weight function (e.g. t-score)

     (the significance of a feature for the given LU)
  2. Features in a row vector of the matrix are sorted in the ascending order on the weighted values.
  3. The k highest-ranking features are selected; e.g. $k = 1000$ works well.
  4. Value of every feature c_i is set to: $k\text{-}ranking(c_i)$

     (a rank according to inverted ranking)
- Cosine similarity measure for rank vectors

Lexico-morphosyntactic Constraints: Verbs

NSb – a particular noun as a potential subject of the given verb

NArg – a noun in a particular case as a potential verb argument

VPart – a present or past participle of the given verb as a modifier of some noun

VAdv – an adverb in close proximity to the given verb

Lexico-morphosyntactic Constraints: Example - Close Adverb (VAdv)

```
or(and(in(pos[0],
        fin,praet,impt,imps,inf,ppas,ppact,pcon,pant),
       llook(-1,begin,$AL,or(
              in(pos[$AL],fin,ger,praet,impt,imps,
                inf,ppas,ppact,pcon,pant,conj,interp),
              and( equal(pos[$AL],adv),
                   inter(base[$AL],"adverb A"))
       )),
       equal(pos[$AL],adv) ),
   and(
       a similar constraint for gerund forms
       and the left context ),
   symmetric constraints for non-gerund verb forms
   and the right context
)
```

Lexico-morphosyntactic Constraints: Adjectives

ANmod – an occurrence of a particular noun as modified by the given adjective

(only nouns which agree on case, gender and number)

AAdv – an adverb in close proximity to the given adjective,

AA – the co-occurrence with an adjective that agrees on case, number and gender

(as a potential co-constituent of the same NP)

- AA was advocated to express negative information

(Hatzivassiloglou and McKeown, 1993)

$$MSR_{Adj}(l_1,l_2) = \alpha\, MSR_{ANmod+AAdv}(l_1,l_2)+\beta\, MSR_{AA}(l_1,l_2)$$

- the best results for: $\alpha = \beta = 0.5$

Experiments: WordNet-Based Synonymy Test

- WordNet-Based Synonymy Test (WBST)
  - claimed to be more difficult than TOEFL used in LSA
  - for a question word q its synonym s is randomly chosen from plWordNet, e.g.

Q: *nakazywać (command)*

A: *polecać (order)*   *pozostawać (remain)*
wkroczyć (enter)   *wykorzystać (utilise)*

Q: *bolesny (painful)*

A: *krytyczny (critical),*   *nieudolny (inept),*
portowy ((of) port),   *poważny (serious)*

Experiments: Data

- The IPI PAN Corpus
  - general Polish, ~254 mln. of tokens
- Verbs
  - 2 984 verbs, 3 086 Q/A pairs in WBST
  - humans (100 Q/A pairs): 88.21% (84-95%)
- Adjectives
  - 2 718 adjectives, 3 532 Q/A pairs in WBST
  - humans (100 Q/A pairs): 88.91% (82-95%)

Experiments: Evaluation for Verbs by WBST

	Frequent LUs				All LUs			
Features	Lin	CRMI	RFF	RWF	Lin	CRMI	RFF	RWF
NArg(acc)	69.60	66.43	56.06	**72.45**	62.56	62.46	45.64	**66.55**
NArg(dat)	**44.97**	19.72	37.53	26.05	**33.58**	17.96	28.65	22.24
NArg(inst)	**64.13**	46.40	49.80	59.07	**52.03**	40.81	41.56	51.02
NArg(loc)	**64.13**	54.47	50.75	62.79	50.18	44.02	39.55	**50.86**
Nsb	62.95	58.35	49.49	**63.18**	51.54	52.38	40.58	**54.94**
VPart	**55.66**	42.04	48.54	46.00	**45.90**	34.94	39.48	41.20
VAdv	72.68	53.60	55.50	**75.30**	62.07	45.67	43.37	**64.02**
Narg(all)	74.82	68.65	56.45	**74.98**	65.51	69.47	46.29	**70.15**
all	76.88	70.23	55.34	**77.12**	68.17	71.99	48.17	**73.45**

- Freitag et. al. (2005): 63.8% for frequent

Experiments: Examples of Verb Lists

ściągnąć (*take off*) [18]

ściągać (*take off* (*habitual*))	0.640
zdjąć (*take off*)	0.608
ubrać (*clothe*)	0.575
założyć (*put on*)	0.562
włożyć (*put on*)	0.554
przyciągnąć (*draw*)	0.552
nosić (*wear*)	0.550
odziać (*clothe*)	0.548
przyciągać (*draw* (*habitual*))	0.542
zrzucić (*drop off*)	0.538

graniczyć (*border*) [8]

sąsiadować (*neighbour*)	0.575
przylegać (*abut*)	0.548,
położyć (*put down*)	0.537
należeć (*belong*)	0.533
zabudować (*build* (*on*))	0.532
zaniedbać (*neglect*)	0.531
dotknąć (*touch*)	0.531
okalać (*encircle*)	0.529
administrować (*administer*)	0.527
otaczać (*surround*)	0.526

Experiments: Examples of a Bad Verb List

okupować (*occupy*) [1]

opuścić (*leave*)	0.556
protestować (*protest*)	0.550
szturmować (*storm*)	0.550
zajmować (*occupy*)	0.543
wyniszczyć (*exterminate*)	0.543
zjednoczyć (*unite*)	0.541
zająć (*occupy*)	0.541
wtargnąć (*invade*)	0.538
maić (*decorate*)	0.537
zabukować (*book*)	0.536

Experiments: Evaluation for Adjectives by WBST

	Frequent LUs				All LUs			
Features	Lin	CRMI	RFF	RWF	Lin	CRMI	RFF	RWF
AAdv	60.05	13.40	62.62	**62.81**	48.65	12.94	49.82	**52.19**
AA	**77.58**	50.47	64.12	76.14	**69.16**	46.30	54.12	68.37
ANmod	**76.39**	71.01	64.06	75.27	71.68	70.60	58.57	**72.47**
Anmod +AAdv	77.40	73.14	65.56	**77.71**	72.25	72.33	59.44	**74.71**
(ANmod+ AAdv)⊕AA	81.65	75.95	67.44	**82.91**	75.70	75.47	61.29	**77.77**
Anmod +AAdv+AA	79.65	76.64	66.12	**79.90**	75.50	76.21	60.52	**77.97**

- Freitag et. al. (2005): 74.6% for frequent

Experiments: Examples of Adjective Lists

niezwykły (*unusual*) [13]

wyjątkowy (*exceptional*)	0.325
niebywały (*unprecedented*)	0.285
niesamowity (*uncanny*)	0.279
niepowtarzalny (*incomparable*)	0.266
wspaniały (*excellent*)	0.250
niespotykany (*unparalleled*)	0.236
niecodzienny (*uncommon*)	0.222
niesłychany (*unheard of*)	0.213
cudowny (*miraculous*)	0.204
szczególny (*particular*)	0.202

agresywny (*aggressive*) [6]

brutalny (brutal)	0.208
odważny (brave)	0.203
dynamiczny (dynamic)	0.189
aktywny (active)	0.189
energiczny (energetic)	0.178
napastliwy (aggressive)	0.176
ostry (sharp)	0.174
arogancki (arrogant)	0.173
wulgarny (vulgar)	0.170
zdecydowany (decided)	0.170

Experiments:
Examples of a Bad Adjective List

kurtuazyjny (*courteous*) [1]

wykrętny (*evasive*)	0.191
kategoryczny (*categorical*)	0.157
oficjalny (*official*)	0.154
urywany (*intermittent*)	0.142
dyskusyjny (*debatable*)	0.139
lakoniczny (*laconic*)	0.138
kawiarniany (*of café*)	0.135
spontaniczny (*spontaneous*)	0.133
retoryczny (*rhetorical*)	0.133
nieoficjalny (*unofficial*)	0.131

MSR and Wordnet Extensions

- Manual assessment of all elements a list
  - $n = 20$, samples with the 95% confidence level
  - positive (head, element) pair: some wordnet relation
  - classes:
    - very useful – a half of the list are positive pairs,
    - useful – a sizable part of the list are positives,
    - neutral – several positives,
    - useless – at most a few positives

PoS	very useful	useful	neutral	useless	no positives
Verb [%]	17.8	37.6	20.0	15.6	9.0
Adjective [%]	19.2	26.3	29.7	14.4	10.4

Observations and future work

- The MSR based on RWF for nouns exhibits comparable performance to MSRs for verbs and adjectives.
- A very small number of morphosyntactic constraints resulted in a relatively high accuracy in the WBST.
  - well above the random baseline in WBST
  - better than reported — though many fewer LUs
  - results closer to human performance than those for nouns
- The method should be easily adapted to similar (similarly inflected) languages, especially Slavic.

Wrocław University of Technology

Corpus-based Semantic Relatedness for the Construction of Polish WordNet

Thank you for your attention

Bartosz Broda[1], Magdalena Derwojedowa[3], Maciej Piasecki[1], Stanisław Szpakowicz[2],

1. Institute of Applied Informatics, WUT
2. Institute of the Polish Language,Warsaw University
3. School of Information Technology and Engineering, University of Ottawa

plwordnet.pwr.wroc.pl