Croatian Science Foundation

Metaphorical collocations

HRZZ (IP-2020-02-6319)

FAKULTET INFORMATIKE I DIGITALNIH TEHNOLOGIJA
SVEUČILIŠTE U RIJECI

A General Framework for Detecting Metaphorical Collocations

M. Brkić Bakarić[1], L. Načinović Prskalo [1], M. Popović [2]

[1] Faculty of Informatics and Digital Technologies, University of Rijeka, Rijeka, Croatia

[2] ADAPT Centre, School of Computing Dublin City University, Ireland

mbrkic@uniri.hr, lnacinovic@uniri.hr, maja.popovic@adaptcentre.ie

Introduction

- aim: to define a framework for detecting metaphorical collocations
- methodology: a combination of computational-linguistic and theoretical-semantic approaches
- languages: Croatian, English, German, Italian
- goal:
  - to explore different patterns involved in the formation of metaphorical collocations in Croatian and discover possibilities of their automatic extraction
  - to create multilingual inventories of metaphorical collocations extracted from comparable corpora
    - universal formation patterns?

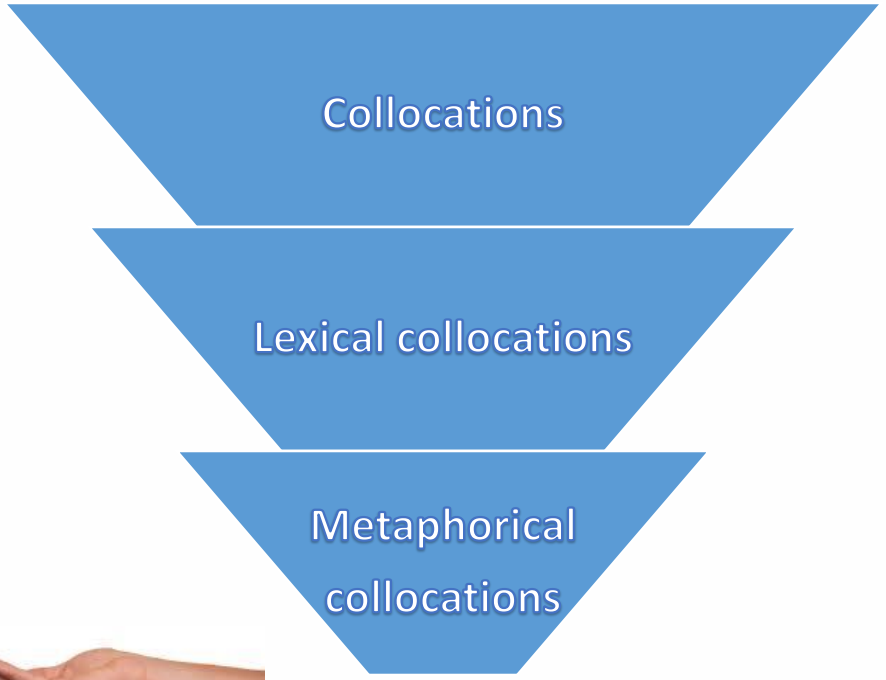
Broadly related work

- rankers (LLR, LDA, SVM, and NN) based on 82 association scores perform better than the individual AMs, although PCA shows that the number of model variables can be significantly reduced (Pecina, 2010)
- a supervised machine-learning approach produces more relevant ranking results than the approach based on heuristics (Ljubešić et al., 2021)
- optimal choice of an AM depends strongly on the particular gold standard used (Evert et al., 2017)
- larger corpora of the same kind perform better (Pecina, 2010; Evert et al., 2017)
- clean, balanced corpora are better than large, messy Web corpora of the same size (Evert et al., 2017)
- word embeddings approach is more useful for ranking than logDice (Ljubešić et al., 2021)
- Croatian
  - linguistic pre-filter + AM (**PMI**, Dice coefficient, χ^2 and LLR) (Petrovic et al., 2006)
  - evolving new AMs shows the importance of POS tags (Šnajder et al., 2008)
  - statistics + linguistic post-filter (Seljan & Gašpar, 2009)
  - decision trees, rule induction, NB, NN, and **SVM (**PMI, semantic relatedness, POS) (Karan, Šnajder and Bašić 2012)
  - Word Sketch + frequency and syntactic post-filter (Hudeček & Mihaljević, 2020)

Metaphorical collocations

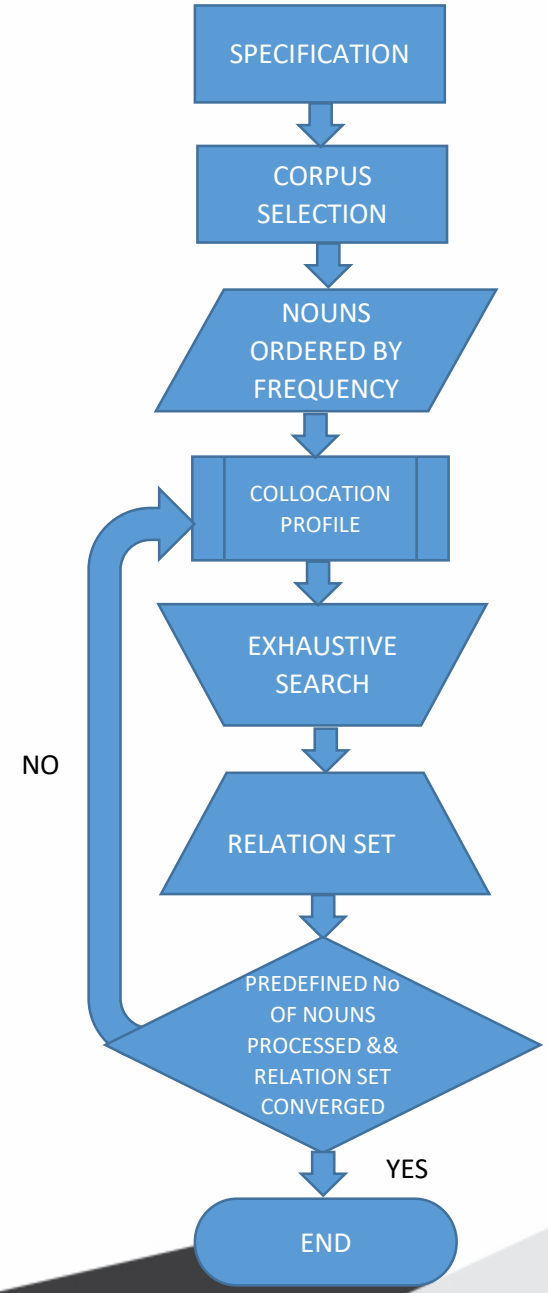
okorjeli neženja

eingefleischther Junggeselle



long-time bachelor

Gold standard compilation framework



	Lemma	Frequency ? ↓
1	godina	4,038,699 •••
2	čovjek	2,276,337 •••
3	dan	2,112,293 •••
4	vrijeme	1,730,336 •••
5	hrvatska	1,503,456 •••

$$logDice(w_1, R, w_2) = 14 + \log_2 \frac{2 \times |(w_1, R, w_2|)}{||w_1, R, *|| + ||*, R, w_2||}$$

COLLOCATION PROFILE

EXHAUSTIVE SEARCH

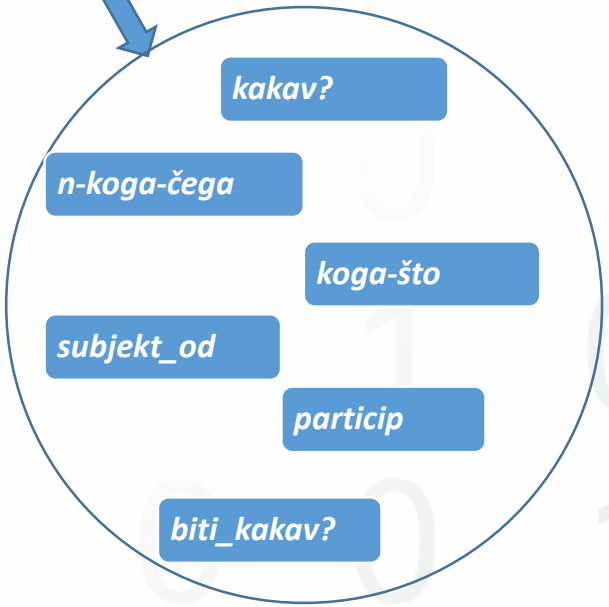
Croatian Web Corpus (Ljubešić & Erjavec, 2011)

	Lemma	Frequency ? ↓
1	time	68,756,908 •••
2	year	64,408,464 •••
3	people	46,762,348 •••
4	day	37,624,460 •••
5	way	35,196,512 •••

	Lemma	Frequency ? ↓
1	jahr	11,283,493 •••
2	zeit	5,294,658 •••
3	mensch	4,696,272 •••
4	tag	4,449,755 •••
5	kind	4,037,595 •••

	Lemma	Frequency ? ↓
1	anno	26,499,692 •••
2	parte	18,298,462 •••
3	tempo	14,259,199 •••
4	giorno	12,461,011 •••
5	volta	11,984,979 •••

kakav	akademski
	cijel
	čitav
n-koga-čega	album
	auto
	dio
subjekt_od	dati
	doći
	ići
particip	ispuniti
	proglasiti
	potrošiti
biti_kakav	rekordan
	sjajan
	rodan
koga_što	ponavljati
	zaključiti
	započeti

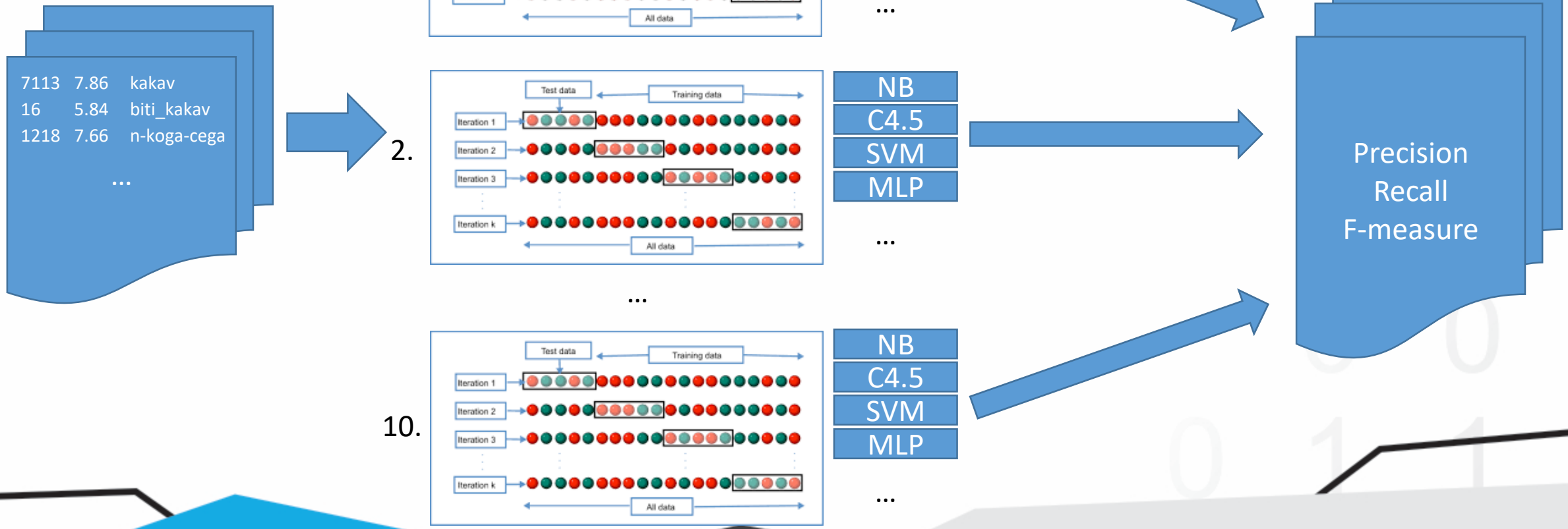


Annotation task results

Relation	# of cands	# of colls	# of m_colls	ratio of m_colls
kakav? (oba_u_genitivu)	99	54	54	55%
n-koga-čega	100	41	38	41%
koga-što	100	41	41	41%
particip	100	16	11	11%
subjekt_od	100	30	30	30%
biti_kakav? +25	74	20	20	55%
Total	673	202	194	29%

- form: collocation – metaphorical collocation – comments
- subtypes:
  - personification over 80% of *subjekt_od*
  - terms slightly superior over metaphors in *n-koga-čega*
  - over 60% of metaphors in *kakav?, koga-što, particip*, and *biti_kakav*
- *N + (N || A || V)*

Experiment


7113 7.86 kakav
16 5.84 biti_kakav
1218 7.66 n-koga-cega
...
1.
2.
10.
Test data
Training data
Iteration 1
Iteration 2
Iteration 3
Iteration k
All data
NB
C4.5
SVM
MLP
...
Precision
Recall
F-measure

Preliminary classification results

Recall	f=3	f=303
NB	**0.94***	0.40
C4.5	0.81	0.79
SVM	0.78	**0.80***
MLP	0.80	0.76

Precision	f=3	f=303
NB	0.68	0.72
C4.5	0.73	0.77
SVM	0.74	0.74
MLP	0.75	**0.78***

f=3 collocation frequency, logDice, and relation
f=303 +collocate word embeddings

F-measure	f=3	f=303
NB	0.78	0.50
C4.5	0.77	0.78
SVM	0.76	0.77
MLP	0.77	0.77

Future evaluation framework

GOLD STANDARD

akademska godina
album godine
godina ispunjena
rekordna godina
ponavljati godinu
…

evaluate

- *Precision*
- *Recall*

Conclusion

- determining a way to encode the relation that refers to collocates contributing the semantic feature to their respective base words
  - procedure for compiling the gold standard of metaphorical collocations is suggested
  - general evaluation framework for our future work is established
  - six significant grammatical relations are determined (the word *godina*)
- collocate embeddings strongly affect the performance of NB
- statistically significant differences are obtained only in precision and recall scores between SVM and MLP with f=303

Thank you! ☺