cjvt Center za jezikovne vire in tehnologije

Frequency vs. semantics

Which is better at ranking collocations?

Nikola Ljubešić

"Jožef Stefan" Institute

Ljubljana, 16.10.2018

arrs

JAVNA AGENCIJA ZA RAZISKOVALNO DEJAVNOST REPUBLIKE SLOVENIJE

KOLOS

Introduction

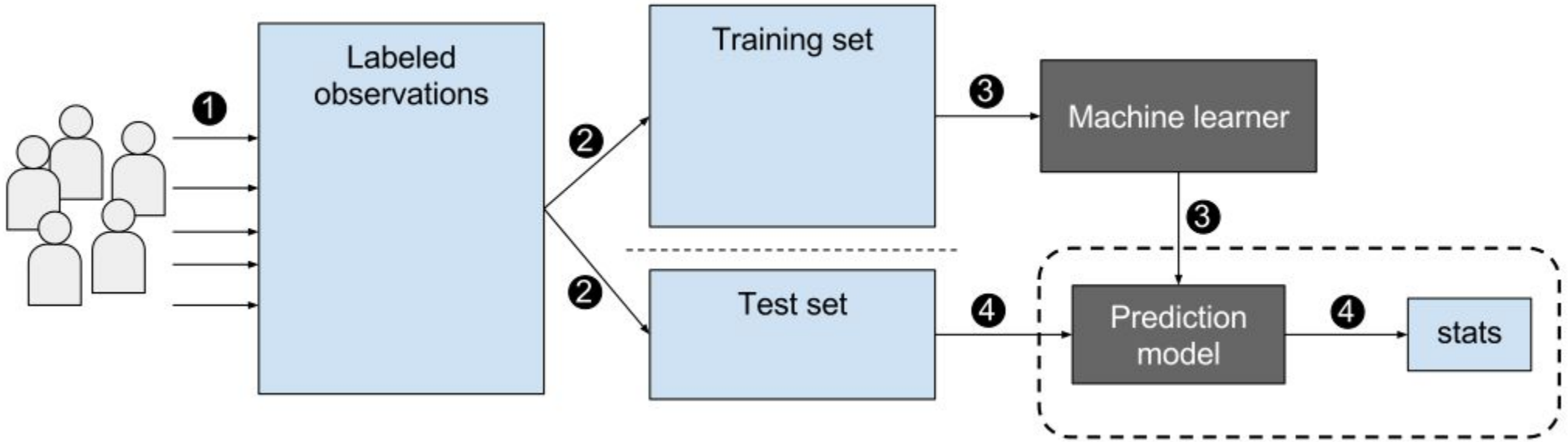
- Ranking collocates by statistical co-occurrence standard approach to enhancing productivity of lexicographers
- Can we improve over this by using supervised machine learning?
- Pecina and Schlesinger (2006) - yes, by combining different association measures (20% relative improvement)
- Our question: can we improve over this by not using association measures, but distributional semantics?
- Broader question: what is more telling for a collocation - frequency or semantics?

Overview

- Background
- Dataset
- Methods
- Results
- Conclusion

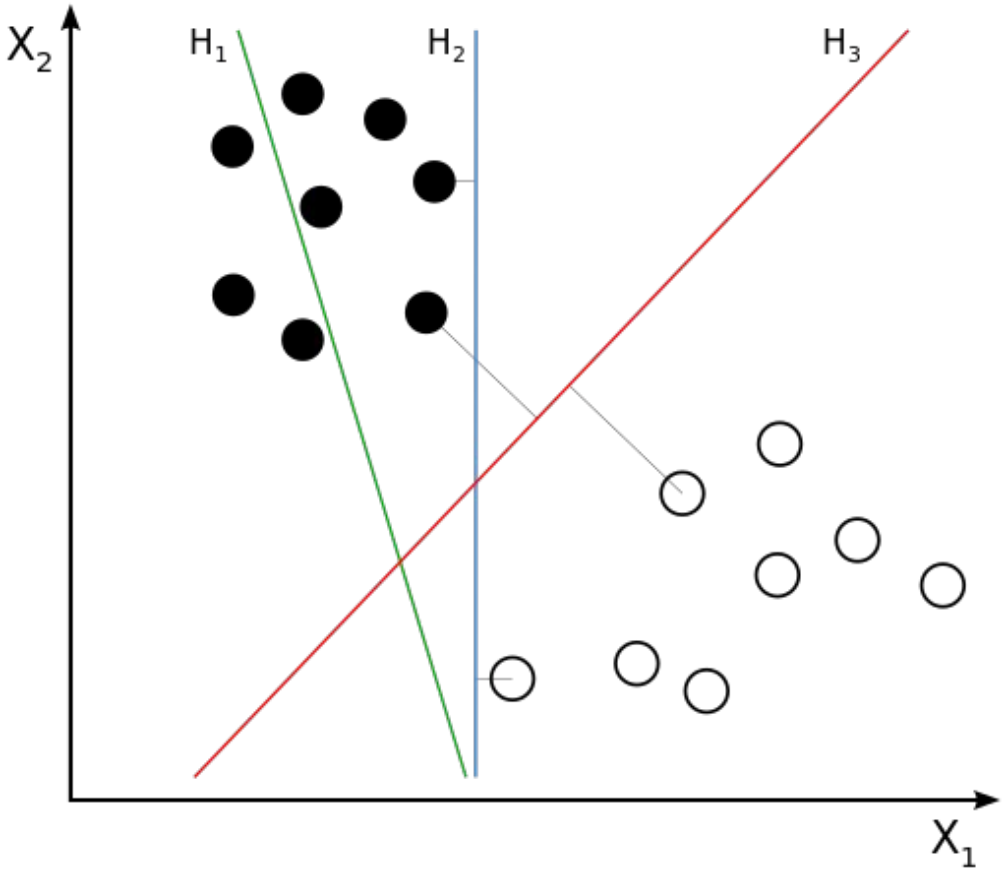
Background

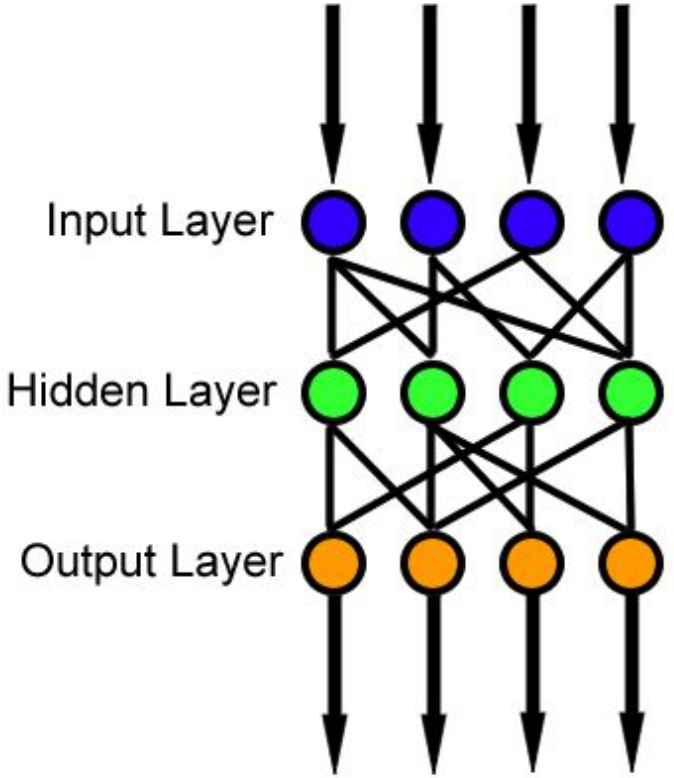
Supervised machine learning



https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/

Support Vector Machine vs. Feed-Forward network

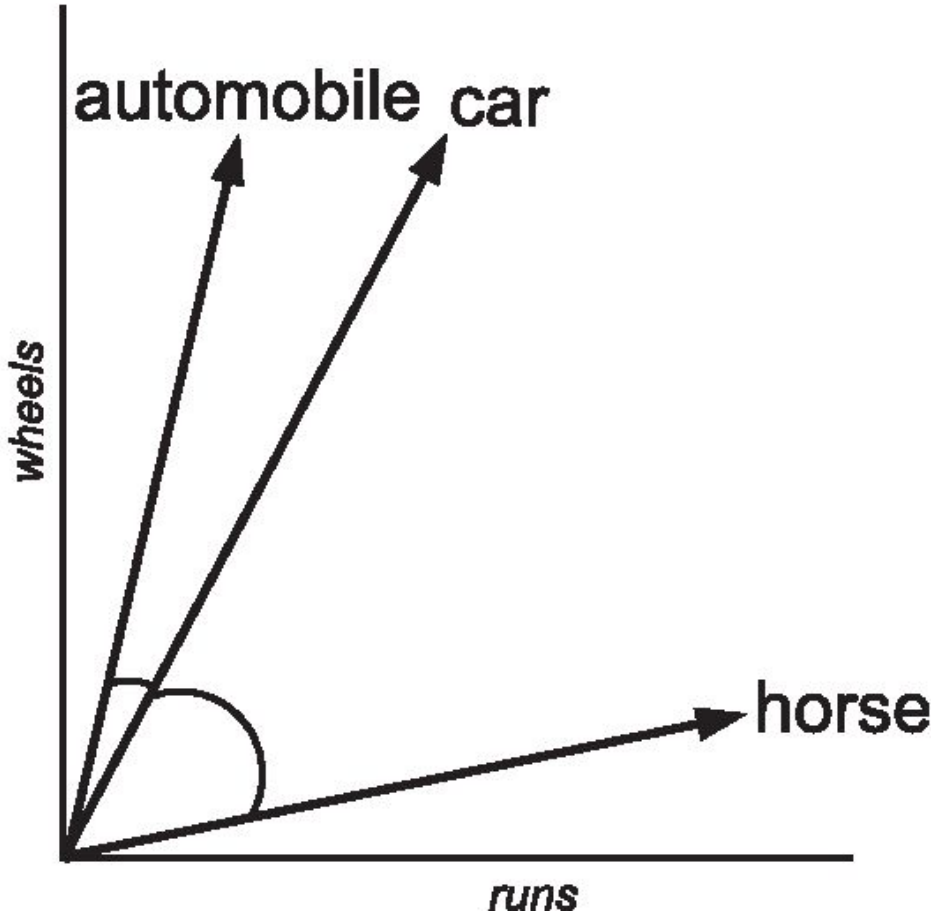




https://en.wikipedia.org/

Distributional semantics

	runs	*wheels*
automobile	1	4
car	2	4
horse	4	1



https://www.semanticscholar.org/paper/Affixation-in-semantic-space%3A-Modeling-morpheme-Marelli-Baroni/8ce36d8c3f356f5e4686de592327665102055be4/figure/0

Distributional semantics via neural networks

- Learning network parameters that maximise the predictiveness of words surrounding a word

[0.20514, -0.38204, -0.43575, -0.35336, -0.19919, -0.1039, 0.067579, -0.12168, 0.67465, -0.30423, -0.25289, 0.047944, 0.48485, -0.24491, 0.30098, -0.11139, -0.45834, -0.36371, -0.049323, -0.36091, -0.26225, -0.25105, 0.29203, -0.059085, -0.066695, -0.29656, 0.54394, -0.0019447, 0.060155, -0.25214, 0.063966, 0.15548, 0.23241, 0.089566, -0.34598, 0.014725, 0.1515, -0.12745, -0.19815, -0.43996, 0.13449, 0.066548, -0.6069, -0.27474, 0.63589, -0.12775, 0.019893, -0.19233, 0.27074, 0.94501, -0.63376, -0.028027, -0.17708, -0.044647, -0.025419, 0.32611, -0.018033, -0.15603, 0.11756, -0.019596, 0.29653, 0.50906, 0.32853, 0.34209, -0.69025, 0.42737, -0.24785, -0.29885, 0.06819, 0.30872, 0.73067, 0.078667, -0.069605, 0.17409, 0.0064074, -0.152, 0.23714, -0.14973, -0.64415, 0.34239, 0.39542, -0.62419, -0.28266, 0.33288, 0.093867, -0.012091, -0.69414, -0.14562, 0.30411, -0.52595, 0.48494, 0.53727, -0.24763, 0.146, 0.23308, -0.48376, -0.07844, 0.71975, -0.10486, -0.69242]

- Parameters for word...
- Nearest neighbours are “seminar”, “symposium”, “give a talk”, “webinar”, “presenter”, “listener” etc.

https://www.semanticscholar.org/paper/Affixation-in-semantic-space%3A-Modeling-morpheme-Marelli-Baroni/8ce36d8c3f356f5e4686de592327665102055be4/figure/0

Dataset

Dataset

- Annotations by five annotators of 17,540 collocation candidates following 130 grammatical relations (gramrels), one final annotation
- Gramrels distributed power-lawish (long-tailed distribution)
- Discard gramrels with less than 20 instances
- 17,142 collocation candidates, 65 gramrels
- Most frequent:
  - pbz0 sbz0, “kisla smetana”, 2594, yes: 2276, no: 318
  - sbz0 sbz2, “brazda pestiča”, 2363, yes: 1931, no: 432
  - gbz sbz4, “segreti žlico”, 1300, yes: 1126, no: 174
  - rbz gbz, “natančno opredeliti”, 1280, yes: 1120, no: 160
  - rbz pbz0, “precej zasoljen”, 765, yes: 486, no: 279
  - sbz0 v sbz5, “satelit v orbiti”, 737, yes: 474, no: 263

Features

Frequency

- Data obtained from the GigaFida corpus via SketchEngine
- **Features** are the following:
  - Headword frequency
  - Collocate frequency
  - Collocation frequency
  - logDice score

Semantics

- Data obtained by learning FastText lemma representations from GigaFida, 100 dimensions
- **Features** are the following
  - 100-dimensional representation of the headword
  - 100-dimensional representation of the collocate
  - By concatenating these representations, we obtain 200 features

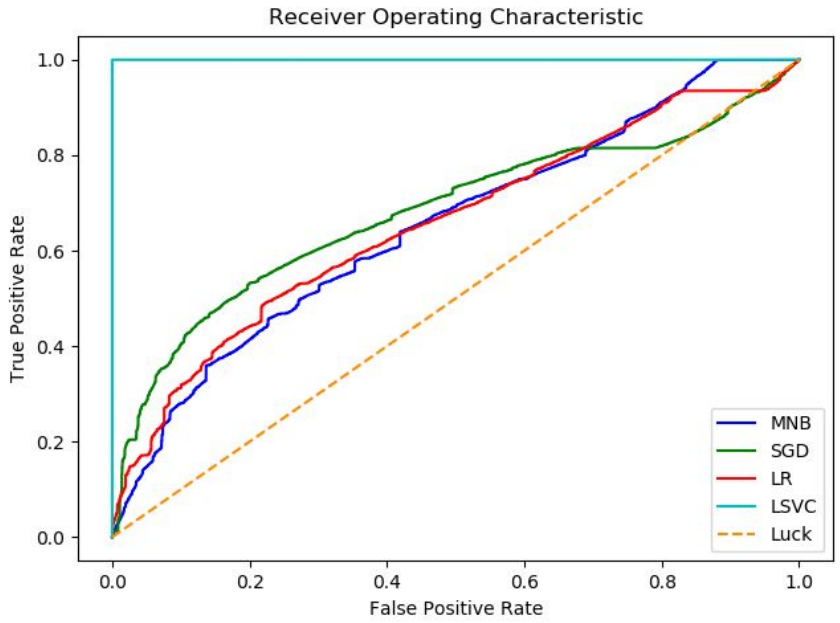
Methods

Systems description

- **logDice** - system using only logDice information, simply ranking candidates by that statistic
- **SkE SVM** - Support Vector Machine (SVM) regressor with scaling, using frequency information (logarithms of frequency) (4 features)
- **sem SVM** - Support Vector Machine (SVM) regressor, using distributional semantic information (200 features)
- **SkE+sem SVM** - SVM using concatenation of frequency and distributional information (204 features)
- **sem FF** - feed-forward neural network, using distributional semantic information (200 features)
- **SkE+sem SVM** - two feed-forward neural networks, encoding separately frequency and distributional information, merging that information in a third feed-forward network (200 and 4 features)

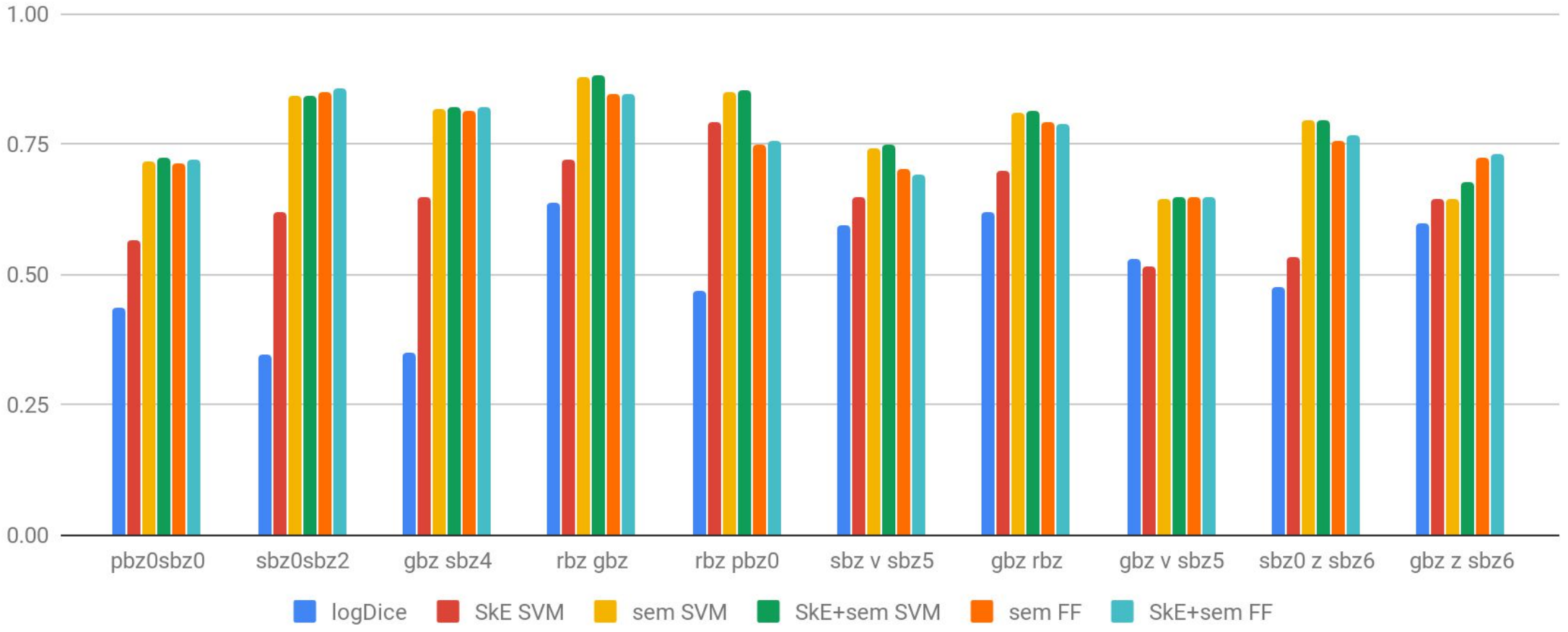
Experimental setup

- Consider the task a ranking task
- Goal - rank positive collocation candidates higher than negative collocation candidates
- Evaluation via Area Under Curve (AUC) score, plot true positive vs. false positive rate and calculate the area below
  - 0.5 if results are random (same proportion of true positives and false positives)
  - 1 if results are perfect, i.e., all true positives higher ranked than any false positives
- Stratified cross-validation with three bins
- Perform separate experiment on each gramrel, merging all gramrels decreases performance



Results

Results by 10 most frequent gramrels



Averaged results

System	Average AUC
logDice	0.488
SkE SVM	0.627
sem SVM	0.738
SkE+sem SVM	0.745
sem FF	0.743
SkE+sem FF	0.744

Initial manual analysis of results

- Compare output of **SkE SVM** and **sem SVM** - isolate the difference in the type of information available for ranking: frequency vs. semantics
- For *rbz gbz* and *rbz pbz0* - order the candidates by difference in ranks of the two systems
- Findings:
  - The semantic approach naturally (over)fits to the lexis available in training data, this is exactly the type of information that we make available to it
  - That approach does not simply memorize lexis, but generalizes as well:
    - Ranks lower temporal, interrogative, modal, conjunctive adverbs, deixis
    - Ranks higher elativ, semantically full adverbs
  - Deeper analyses needed to identify potential interaction between representations of the headword and the collocate

Conclusion

Conclusion

- logDice incapable of ranking properly the top of the list (potential issue - rankings are merged!, evaluating separately rankings and averaging?)
- Frequency information useful in a supervised setting (AUC of 0.63)
- Semantic information much more potent (AUC of 0.74)
- Merging the two sources of information improves the results slightly (2.7% relative error reduction on SVM, less on FF)
- SVM as good as FF - surprise as FF should be much better at handling variable interactions
- Next steps
  - Deeper analysis of the differences in the results (both linguistic and technical)
  - Merging similar gramrel instances - those that only differ in the preposition?
  - Overrepresenting positive instances from collocation dictionaries

cjvt Center za jezikovne vire in tehnologije

Frequency vs. semantics

Which is better at ranking collocations?

Nikola Ljubešić

"Jožef Stefan" Institute

Ljubljana, 16.10.2018

arrs

JAVNA AGENCIJA ZA RAZISKOVALNO DEJAVNOST REPUBLIKE SLOVENIJE

KOLOS