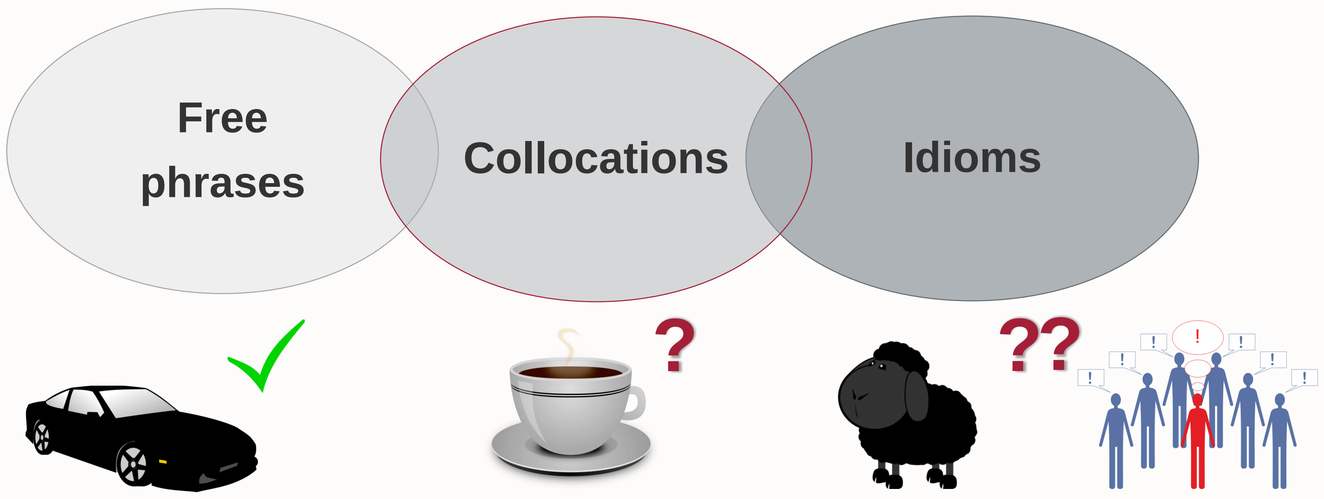
Automatic Collocation Identification Using Word Embeddings

Yana Strakatova
University of Tübingen



Collocations

- syntactic relation
- partial semantic opacity
- one of the elements carries a special meaning found only in this combination
- constrained lexical selection
- high degree of statistical association

Association Measures (AMs)

✓ extract frequency counts for the words from a corpus
✓ create ranked n-best lists according to the association measure scores

	f_1 adjective	f_2 noun	f adj+noun	Association (dice)
red army	880	160	22	**0.04**
red carpet	880	44	15	0.03
red rose	880	46	8	0.01
red dress	880	84	2	0.004

Word Embeddings

n-dimensional, real-valued vectors
co-occurrence frequencies of words (context):

context ->	bitter	black	fast	**vector**
coffee	2	7	0	[2,7,0]
tea	1	5	0	[1,5,0]
boat	0	1	6	[0,1,6]

He loved his *black* **coffee**, crackers, and an orange.

Dataset (Evert 2008):

- German adj+noun lemma combinations
- annotated by lexicographers
- unique adjectives: 489
  unique nouns: 815

520 true collocations (categories 1&2): *groß+Liebe* 'great love', *grün+Politik* 'green politics'

732 non-collocations: *groß+Park* 'big park', *grün+Baum* 'green tree'

Collocation extraction: previous experiments

best AMs (Evert 2004, 2008):

1. dice
2. log-likelihood
3. chi-squared

best AMs and combining multiple AMs (Pecina 2008):

Method	Average Precision
Piatersky–Shapiro coefficient	**0.63**
LDA (multiple AMs)	0.61
GLM (multiple AMs)	0.60

Binary classification of collocation candidates

Input: **word embeddings** (Dima 2015) **association scores**

brennendes 'burning' + *Problem* 'problem' = **brennendes Problem** + dice, log-likelihood, chi-squared, MI

Logistic Regression binary classifier

- 80% training-20% test
- 10-fold cross-validation

Output: classified candidates
1 – a true collocation
0 – not a collocation

Evaluation

Features	Average Precision	f1-score
baseline (random)	0.42	0.46
300 dimensions	0.71	0.68
50 dimensions + AMs	0.69	0.70
100 dimensions + AMs	0.71	0.69
200 dimensions + AMs	0.71	0.69
300 dimensions + AMs	**0.72**	**0.70**

REFERENCES
Dima, C. (2015). Reverse-engineering language: A study on the semantic compositionality of german compounds. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1637–1642, Lisbon, Portugal. Association for Computational Linguistics.
Evert, S. (2004). The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.
Evert, S. (2008). A lexicographic evaluation of german adjective-noun collocations. Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), Marrakech, Morocco.
Pecina, P. (2008). Lexical Association Measures: Collocation Extraction. PhD thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic.