Do multi-sense word embeddings learn more senses?

Márton Makrai

K + K = 120 Workshop 2017

Overview

Word embeddings

Multi-sense

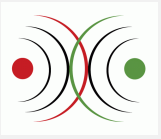
Experiments

Overview

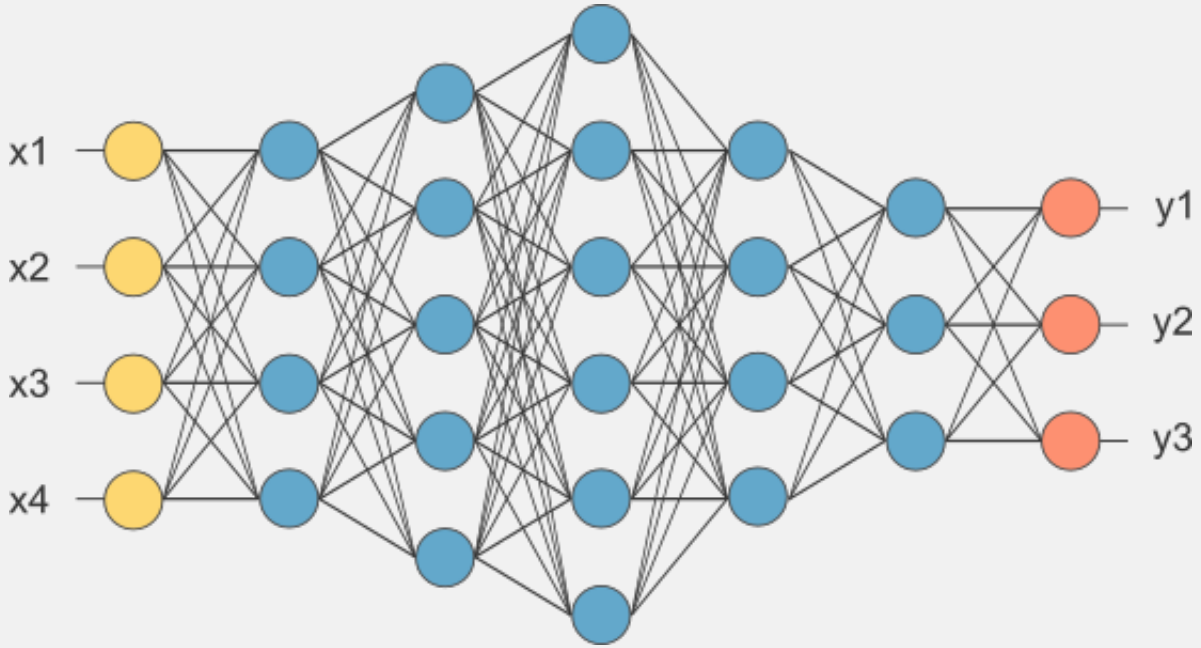
Word embeddings

Multi-sense

Experiments



Artificial neural networks



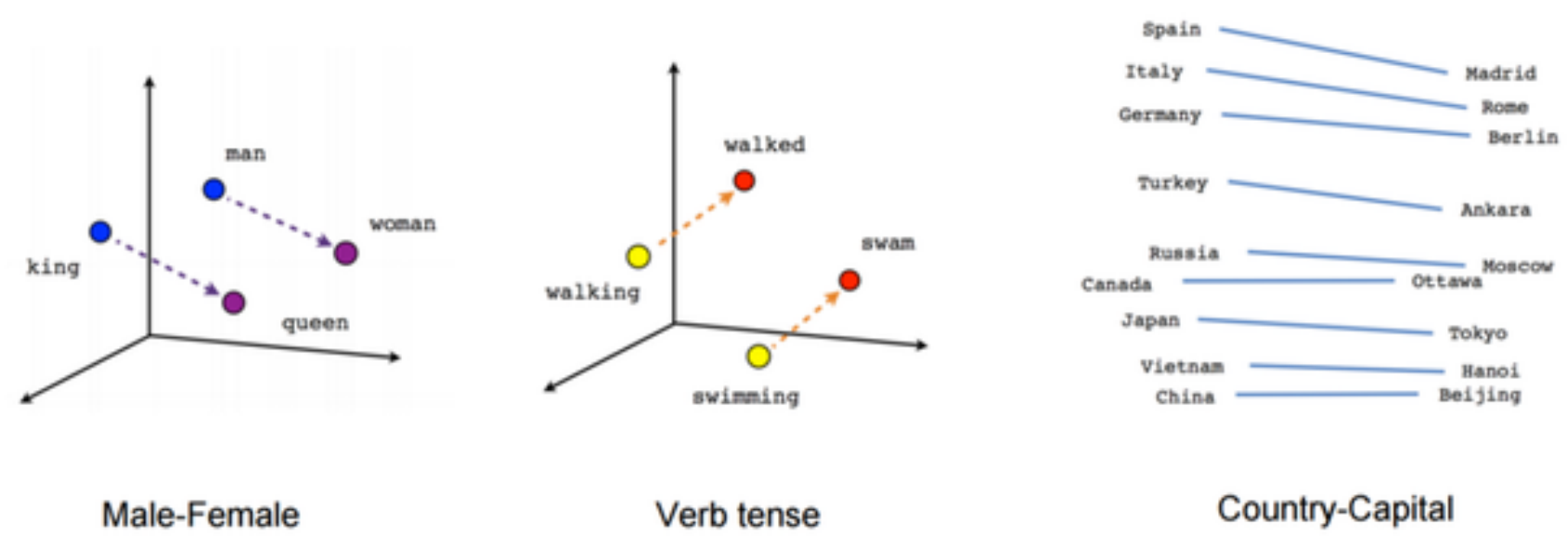
- cybernetics (1949), connectionism (1974), deep learning (2006)
- Learning features, more and more abstract layers
- - computer vision (Krizhevsky and Sutskever, 2012)
  - speech recognition (Hinton et al., 2012)
- fast learning on the graphics card
- like in the brain?

Word embeddings

- Representation of words in neural networks
- $\mathbf{w} \in \mathbb{R}^{300}$
- words with similar distribution $\rightsquigarrow$ similar points
- unsupervised training on giga-word corpora
- word2vec: skip-gram or continuous bag of words (Mikolov et al., 2013a)
- representation sharing (Collobert et al., 2011; Hashimoto et al., 2017)
- compositionality
  character, morph, word, query, sentence, rhetorics
  - morphs (Lazaridou et al., 2013)
  - bellow the word level: fastText (Bojanowski et al., 2016)
  - thought vector (Vaswani et al., 2017)

Meaning decomposition with vectors

Katz and Fodor, 1963; Mikolov, Yih, and Zweig, 2013



$$\textbf{king} + \textbf{woman} - \textbf{man} \approx \textbf{queen}$$

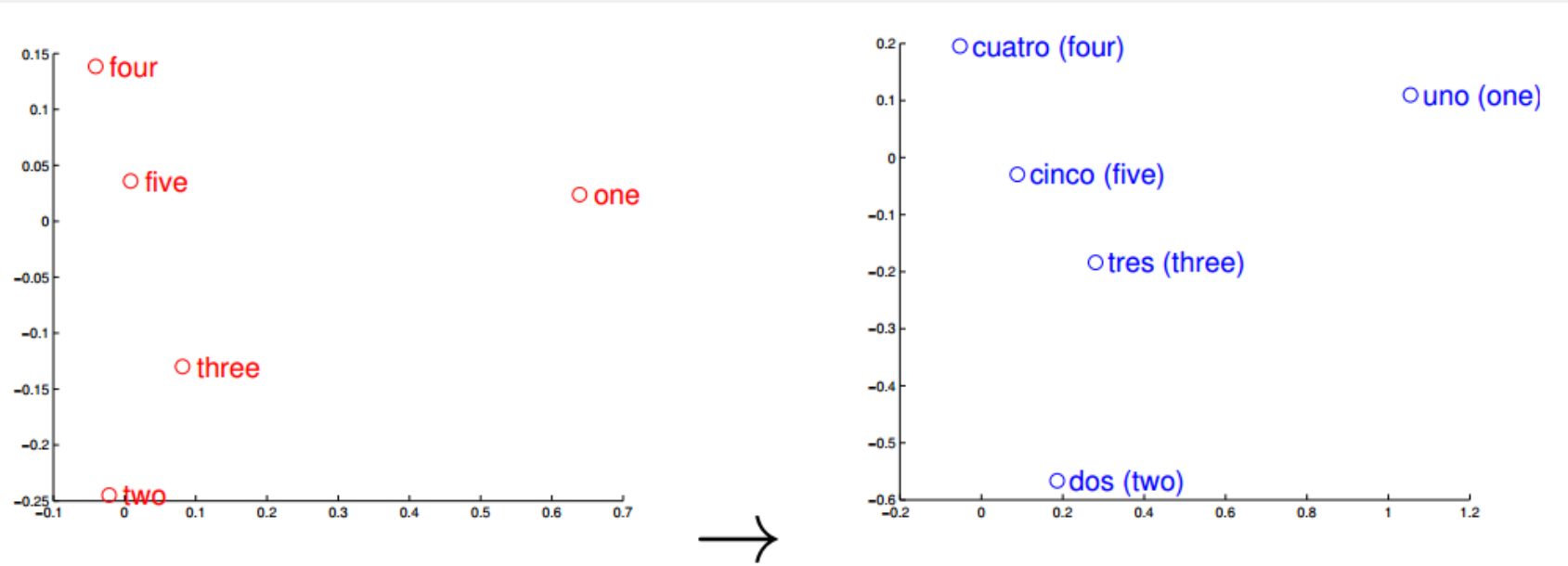
- nearest neighbors

Word translation (Mikolov et al., 2013)

- linear mapping between embeddings, 600 → 300 dim
- training on the 5 000 most frequent pairs
- test on the next 1 000

$$W : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2} \quad z \approx Wx$$

$$\min_W \sum_i ||Wx_i - z_i||^2$$

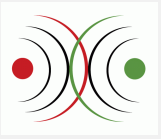


Overview

Word embeddings

Multi-sense

Experiments



Word ambiguity

- homonymy: Russian *mir* ‘world’; ‘peace’
- polysemy: Hungarian *nap* ‘Sun; day’
- evidence for differentiation
  - etymology: common origin
    - uncertain for many words
    - how far back?
  - relatedness of meanings (intuition. Agreement?)

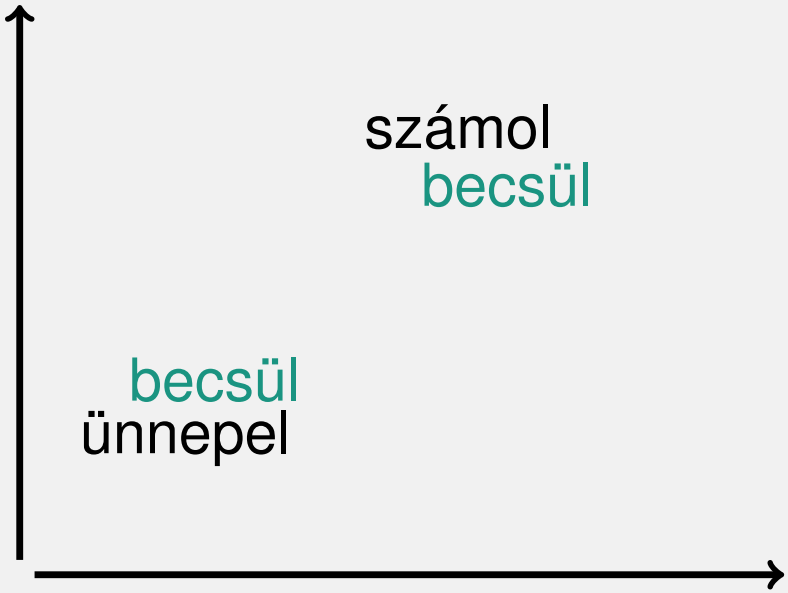
disambiguation (WSD)	classification	supervised
induction (WSI) (Schütze, 1998)	clustering	unsupervised

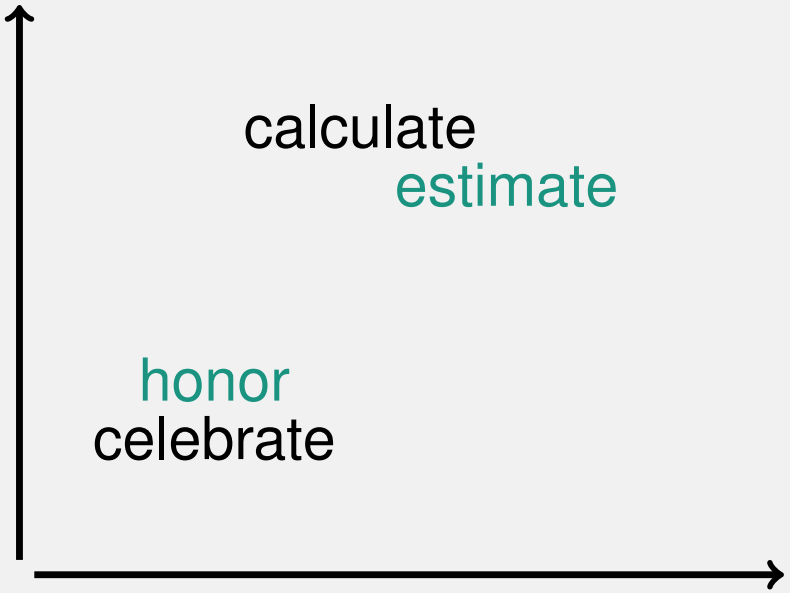
Multi-sense vector models

- multi-“prototype” embeddings (Reisinger and Mooney, 2010)
- with neural network (Huang et al., 2012)
- multi-sense skip gram, open-source tools
  - Neelakantan et al., 2014
  - as a Dirichlet Process
    - `AdaGram` (Bartunov et al., 2016)
    - `mutli` (Li and Jurafsky, 2015)
- if the number of parameters is controlled (Li and Jurafsky, 2015)
  - slight performance boost in
    - semantic similarity for words and sentences,
    - semantic relation identification,
    - part-of-speech tagging
  - no improvement in
    - sentiment analysis
    - named entity extraction
- sense resolution is too fine (duplicates)

Linear translation from multi-sense embedding

- Borbély, Makrai, Nemeskey, and Kornai 2016
- principle: homonymous senses $\rightsquigarrow$ different translations
- target embedding remains single-sense
  (Pennington, Socher, and Manning, 2014; Mikolov et al., 2013b)





Overview

Word embeddings

Multi-sense

Experiments

Data

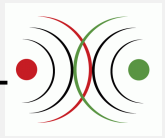
- source corpus
  - de-glutinized version (Borbély et al., 2016a; Nemeskey, 2017) of the Hungarian National Corpus (Oravecz, Váradi, and Sass, 2014)

`jelmondatával`	→	`jelmondat <POSS> <CAS<INS»`	‘(with its) motto’
`akartak`	→	`akar <PAST> <PLUR>`	‘(they) want(ed)’

- target embedding: GloVe 840B 300d (Pennington, Socher, and Manning, 2014)
- seed dictionary: wikt2dict (Ács, Pajkossy, and Kornai, 2013)
- training on the first meaning

Examples

	sim		covg	
S	0.0974	kapcsolat	affair, conjunction, linkage	0.33
S	0.136	futó	runner, bishop	1.0
I	0.1361	kocsi	coach, carriage	1.0
S	0.1626	fogad	bet, greet	1.0
S	0.1873	induló	march, candidate	1.0
S	0.2052	zavar	disturbance, annoy, disturb, turmoil	0.57
S	0.2206	bemutató	exhibition, presenter	0.67
I	0.2494	gazda	farmer, boss	0.67
I	0.2506	kapu	gate, portal	1.0
I	0.2515	előbbi	anterior, preceding	0.67
I	0.2558	kötelezettség	engagement, obligation	0.67
S	0.2807	sorozat	suite, serial, succession	1.0
S	0.2935	durva	coarse, gross	0.18
I	0.3097	megkülönböztetés	discrimination, differentiation	0.5
I	0.319	hirdet	advertise, proclaim	1.0
I	0.3299	aláírás	signing, signature	0.67



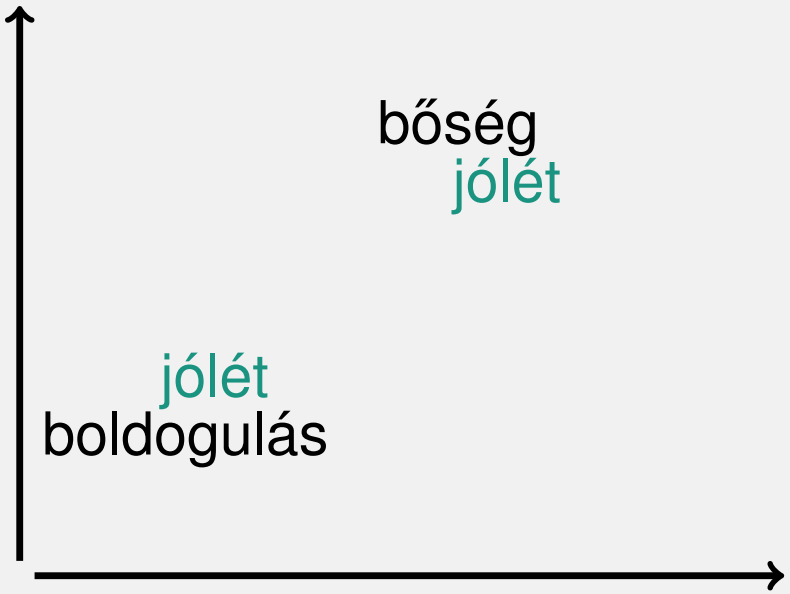
The resolution trade-off

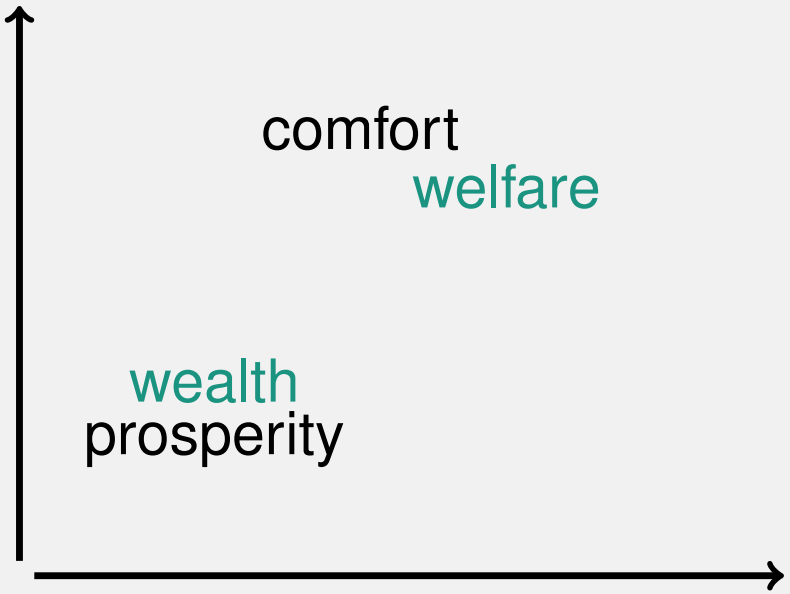
- different vectors $\overset{?}{\Rightarrow}$ different meanings
- `lax`: selection of parameters and target embedding
  - at least one meaning vector should have a good translation
- `disamb`: different sense vectors should have a different set of good translations
  - ratio of such items among those predicted to be ambiguous

	`lax`	`disamb`
`AdaGram`	73.3%	18.53%
`mutli` "sense vectors"	71.0%	19.46%
`mutli` "context vectors"	69.9%	**20.76%**

$$p(w_i \mid w_j) \propto \exp(u_i^\top v_j)$$

Problem: synonymous translations





Happy Birthday!

Some of the most ambiguous 25 words

		sim		covg
mutli “context vs”	sokaság	0.07848	plurality crowd multitude	0.38
	kar	0.1008	arm choir	1.0
	alkalmazás	0.1087	adaptation hiring employ app	0.67
	bejelent	0.1119	announce lodge	1.0
	csomó	0.116	lump mat knot	1.0
	összeállítás	0.1247	binding compilation editing composition	0.8
	agy	0.1746	butt hub	1.0
	találkozó	0.1898	reunion appointment	1.0
AdaGram	fordítás	0.06056	turning compilation translation	0.75
	ruha	0.1154	dress costume rig clothes garment	0.62
	alkalmazás	0.1236	app employ	0.33
	törzs	0.1308	tribe stem trunk waist hull	0.62
	függő	0.145	dependent aerial addict	0.6
	hangsúlyoz	0.1595	stress accent	0.67
	nyom	0.2582	clue squeeze weigh hint push trace slot foil	0.62
	mag	0.2634	kernel seed	0.4

Bibliography I


Ács, Judit, Katalin Pajkossy, and András Kornai (2013). "Building basic vocabulary across 40 languages". In: *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 52–58 (cit. on p. 13).


Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (2016). "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.


Bartunov, Sergey et al. (2016). "Breaking Sticks and Ambiguities with Adaptive Skip-gram". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)* (cit. on pp. 10, 24).


Bojanowski, Piotr et al. (2016). "Enriching Word Vectors with Subword Information". In: *arXiv preprint arXiv:1607.04606* (cit. on p. 5).


Borbély, Gábor et al. (2016a). "Denoising composition in distributional semantics". In: *DSALT: Distributional Semantics and Linguistic Theory*. poster (cit. on p. 13).

Bibliography II

- Borbély, Gábor et al. (2016b). “Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation”. In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Berlin, Germany: Association for Computational Linguistics, pp. 83–89. DOI: 10.18653/v1/W16-2515. URL: http://www.aclweb.org/anthology/W16-2515 (cit. on p. 11).
- Collobert, R. et al. (2011). “Natural Language Processing (Almost) from Scratch”. In: *Journal of Machine Learning Research (JMLR)* (cit. on p. 5).
- Dinu, Georgiana, Angeliki Lazaridou, and Marco Baroni (2015). “Improving Zero-shot Learning by Mitigating the Hubness Problem”. In: *ICLR 2015, Workshop Track* (cit. on p. 27).
- Faruqui, Manaal and Chris Dyer (2014). “Improving vector space word representations using multilingual correlation”. In: *EACL*. Association for Computational Linguistics, pp. 462–471.
- Hashimoto, Kazuma et al. (2017). “A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks”. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (cit. on p. 5).
- Hinton, G. et al. (2012). “Deep neural networks for acoustic modeling in speech recognition”. In: *IEEE Signal Processing Magazine* 29, pp. 82–97 (cit. on p. 4).

Bibliography III

- Huang, Eric et al. (2012). “Improving Word Representations via Global Context and Multiple Word Prototypes”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*. Jeju Island, Korea: Association for Computational Linguistics, pp. 873–882 (cit. on p. 10).
- Katz, J. and Jerry A. Fodor (1963). “The structure of a semantic theory”. In: *Language* 39, pp. 170–210 (cit. on p. 6).
- Korn, F. and S. Muthukrishnan (2000). “Influence sets based on reverse nearest neighbor queries”. In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (cit. on p. 27).
- Krizhevsky, A. and G. Sutskever I.and Hinton (2012). “ImageNet classification with deep convolutional neural networks”. In: *NIPS’2012* (cit. on p. 4).
- Lazaridou, Angeliki, Georgiana Dinu, and Marco Baroni (2015). “Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning”. In: *ACL*. Long, Oral (cit. on p. 27).
- Lazaridou, Angeliki et al. (2013). “Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics”. In: *ACL (1)*, pp. 1517–1526. URL: `http://aclweb.org/anthology/P/P13/P13-1149.pdf` (cit. on p. 5).

Bibliography IV

- Li, Jiwei and Dan Jurafsky (2015). "Do Multi-Sense Embeddings Improve Natural Language Understanding?" In: *EMNLP* (cit. on pp. 10, 25).
- Mikolov, Tomas, Quoc V Le, and Ilya Sutskever (2013). "Exploiting similarities among languages for machine translation". Xiv preprint arXiv:1309.4168 (cit. on p. 7).
- Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic Regularities in Continuous Space Word Representations". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*. Atlanta, Georgia: Association for Computational Linguistics, pp. 746–751 (cit. on p. 6).
- Mikolov, Tomas et al. (2013a). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems 26*. Ed. by C.J.C. Burges et al. Curran Associates, Inc., pp. 3111–3119. URL: `http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf` (cit. on p. 5).
- Mikolov, Tomas et al. (2013b). "Efficient Estimation of Word Representations in Vector Space". In: *Proceedings of the ICLR 2013*. Ed. by Y. Bengio and Y. LeCun (cit. on p. 11).

Bibliography V

- Neelakantan, Arvind et al. (2014). “Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space”. In: *EMNLP* (cit. on p. 10).
- Nemeskey, Dávid Márk (2017). “`emMorph` a Hungarian Language Modeling baseline”. In: *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*. Szeged, pp. 91–102. arXiv: `1701.07880 [cs.CL]` (cit. on p. 13).
- Oravecz, Csaba, Tamás Váradi, and Bálint Sass (2014). “The Hungarian Gigaword Corpus”. In: *Proceedings of LREC 2014* (cit. on p. 13).
- Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). “Glove: Global Vectors for Word Representation”. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)* (cit. on pp. 11, 13).
- Radovanović, M, A Nanopoulos, and M Ivanović (2010). “Hubs in space: Popular nearest neighbors in high-dimensional data”. In: *Journal of Machine Learning Research* 11, pp. 2487–2531 (cit. on p. 27).
- Reisinger, Joseph and Raymond J Mooney (2010). “Multi-prototype vector-space models of word meaning”. In: *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 109–117 (cit. on p. 10).

Bibliography VI

- Schütze, Hinrich (1998). “Automatic word sense discrimination”. In: *Computational linguistics* (cit. on p. 9).
- Singh, Amit, Hakan Ferhatosmanoğlu, and Ali Şaman Tosun (2003). “High dimensional reverse nearest neighbor queries”. In: *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM)*.
- Vaswani, Ashish et al. (2017). “Attention is All You Need”. In: *NIPS*. URL: `https://arxiv.org/pdf/1706.03762.pdf` (cit. on p. 5).
- Xing, Chao et al. (2015). “Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation”. In: *NAACL*, pp. 1005–1010.
- Youn, Hyejin et al. (2016). “On the universal structure of human lexical semantics”. In: *PNAS* 113.7, pp. 1766–1771.

Number of senses against frequency I

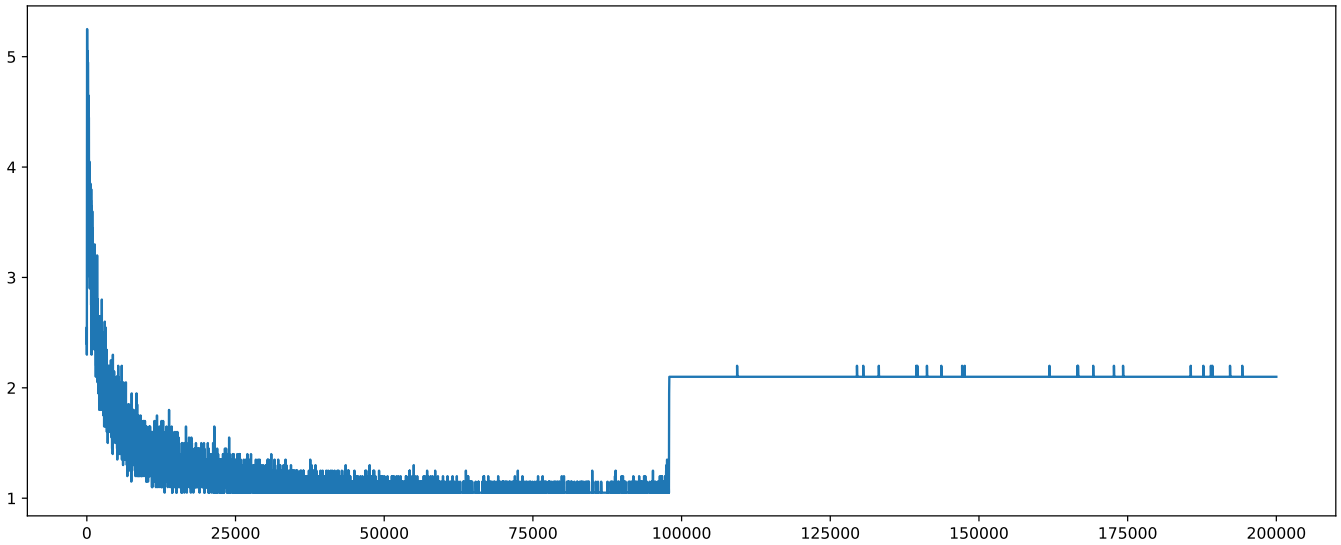


Figure: `AdaGram` (Bartunov et al., 2016)

Ambiguity jumps at frequency 90.

Number of senses against frequency II

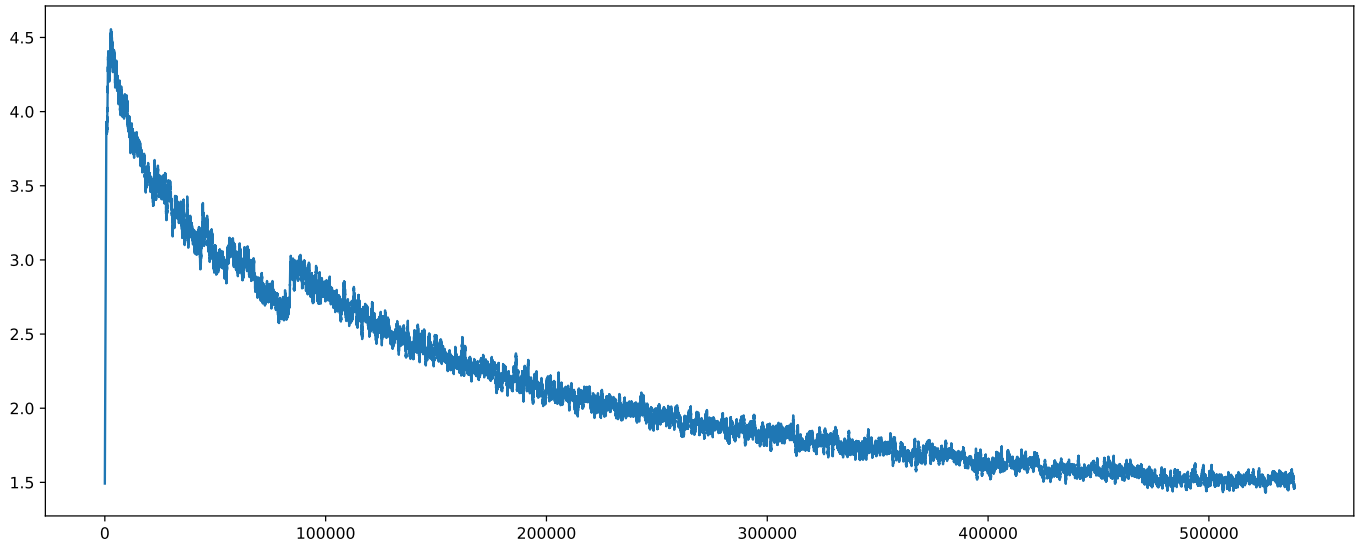


Figure: `mutli`, Chinese Restaurant Process (Li and Jurafsky, 2015)

Number of senses against frequency III

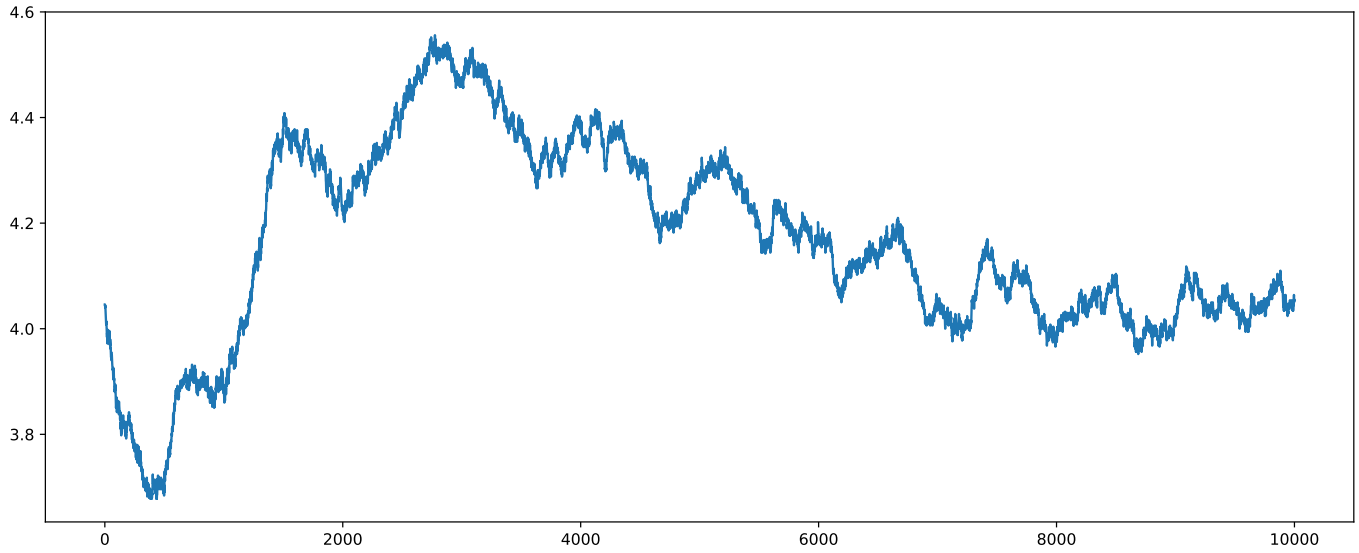


Figure: `mutli`, top 10 K word senses

Reverse nearest neighbors

- nearest neighbor (NN) search in high-dimensional spaces $\rightsquigarrow$
- *hubs*, data points that are NNs of many points
  - wrong in most of the cases (Radovanović, Nanopoulos, and Ivanović, 2010)
- solution: reverse nearest neighbor (revNN) queries (Korn and Muthukrishnan, 2000)
- reverse neighborhood rank
  - *project* is a kth revNN of *work* $\overset{\Delta}{\Leftrightarrow}$ *work* is the kth NN of *project*
- there may be zero, one, or more kth revNNs of a word
- revNN query: return the words with the lowest revNN ranks
- less prone to hubs
- in linear translation (Dinu, Lazaridou, and Baroni, 2015; Lazaridou, Dinu, and Baroni, 2015)