Do multi-sense word embeddings learn more senses?

Márton Makrai
MTA Research Institute of Linguistics

math.bme, 2020. május 18.

Overview

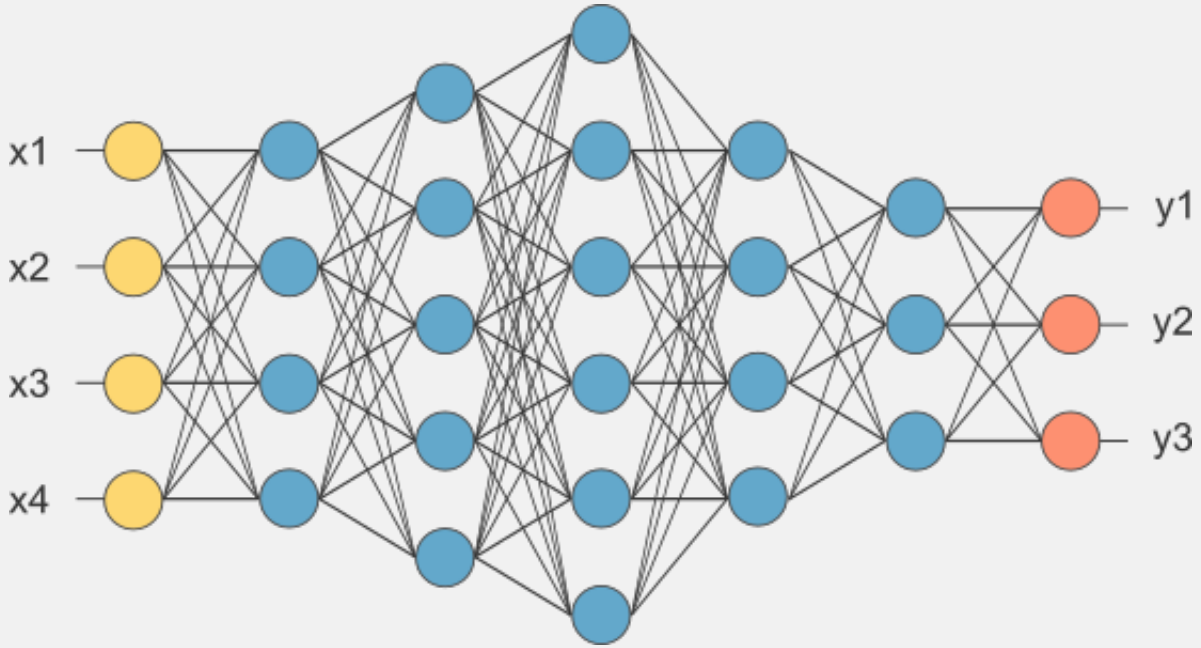
Word embeddings
- Static
- Contextualized

Word senses
- Results

Overview

- Word embeddings
  - Static
  - Contextualized
- Word senses
  - Results

Artificial neural networks



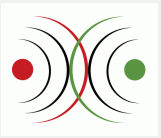
- neural i.e. like the brain?
- cybernetics (1949), connectionism (1974), deep learning (2006)
- Learning features, more and more abstract layers
- - computer vision (Krizhevsky and Sutskever, 2012)
  - speech recognition (Hinton et al., 2012)
  - language (Peters et al., 2018a; Devlin et al., 2018)
- architectures + data + hardware + pretraining

Overview

- Word embeddings
  - Static
  - Contextualized
- Word senses
  - Results

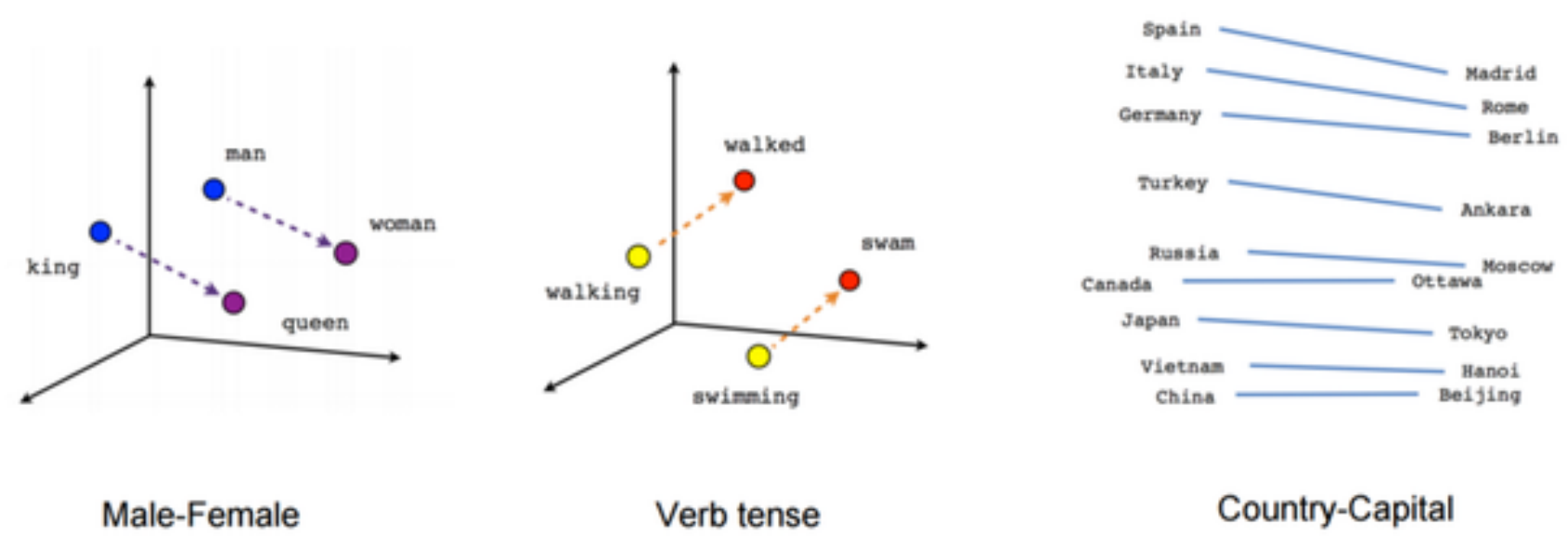
Word embeddings (static)

- before Peters et al., (2018a), in a galaxy...
- 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)
- linguistic levels:
  character $<$ morph $<$ word $<$ query $<$ sentence $<$ rhetorics
  - morphs (Lazaridou et al., 2013)
  - bellow the word level: fastText
    - (Bojanowski, Joulin, and Mikolov, 2016)
  - thought vector (Vaswani et al., 2017)



Meaning decomposition with vectors

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



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

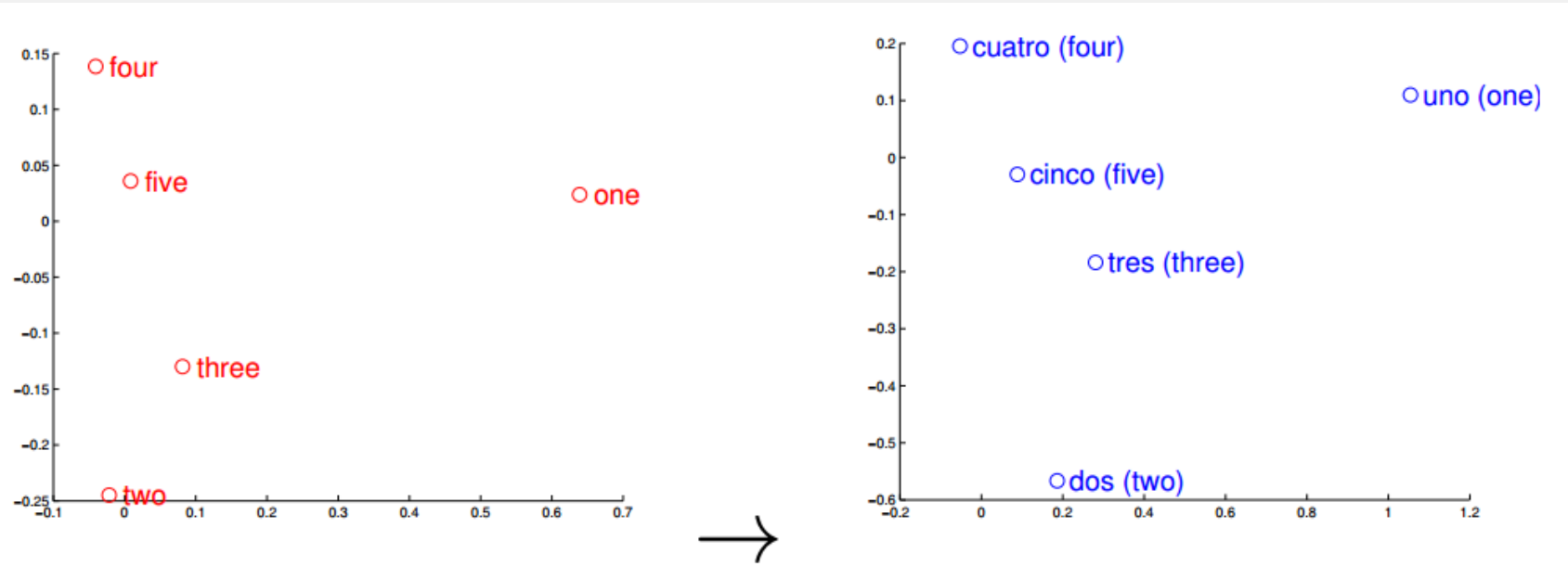
- nearest neighbors

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

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

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

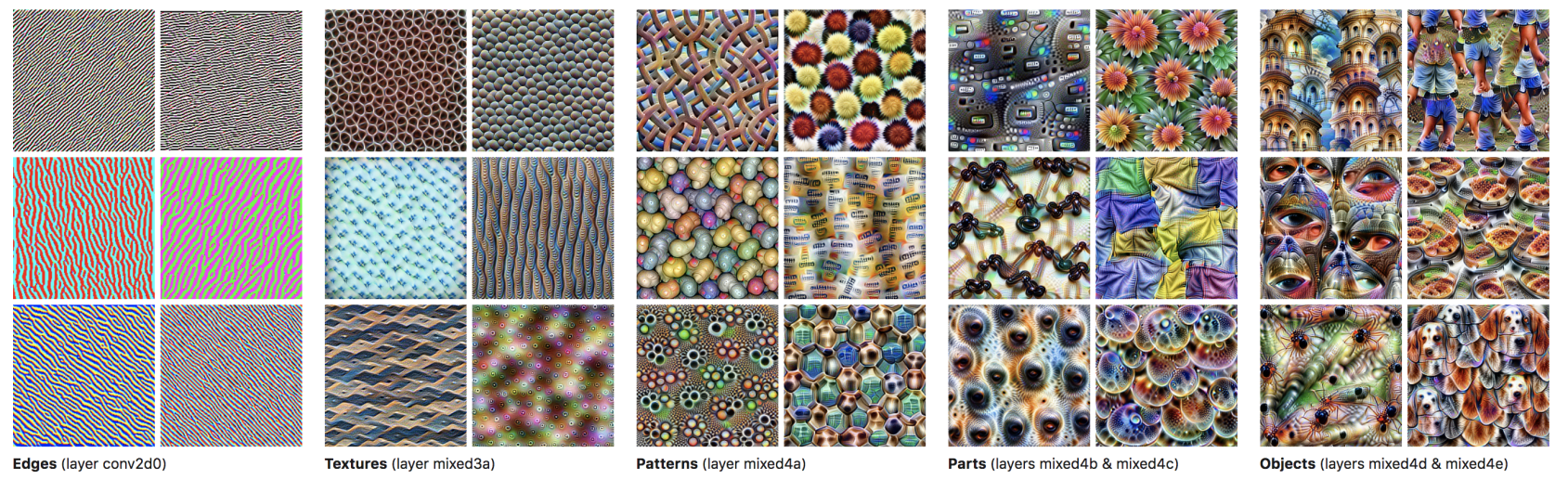


Overview

- Word embeddings
  - Static
  - Contextualized
- Word senses
  - Results

Pre-trained deep language models
aka. contextualized word representations

- Peters et al., 2018a; Devlin et al., 2018
- NLP's ImageNet moment, `https://ruder.io/nlp-imagenet/`
- vision: edges → textures → patterns → parts → objects
- language: morphology → syntax → semantics (and pragatics)



- multilingual transfer

Interpretability in deep language models I
Belinkov and Glass, 2019; Coenen et al., 2019

- accountability, trust, fairness, safety, and reliability
- earlier NLP work: feature-engineering
  - features: morphological properties, lexical classes, syntactic categories, semantic relations
  - one could observe the importance of each feature
    - at least in theory
- approaches to probing models (Conneau et al., 2018)
  - visualization
  - adversarial examples
- representations by layer (Peters et al., 2018b)
- individual neurons in a network may have meaning (Dalvi et al., 2019)
- syntax trees represented in a neural network's word representation space (Hewitt and Manning, 2019)

Interpretability in deep language models II
Belinkov and Glass, 2019; Coenen et al., 2019

- rich morphology (Şahin et al., 2019)
  - word-level probing tasks
    - case marking, possession, word length, morph tag count and pseudo-word identification, 24 languages
  - they relate the probing task performance to classic NLP tasks
    - POS-tagging, dep parsing, sem role labeling, NER, and nat lang infer
- finite or infinite number of context-specific representations/word-senses? (Ethayarajh, 2019)

Overview

Word embeddings
- Static
- Contextualized

Word senses
- Results

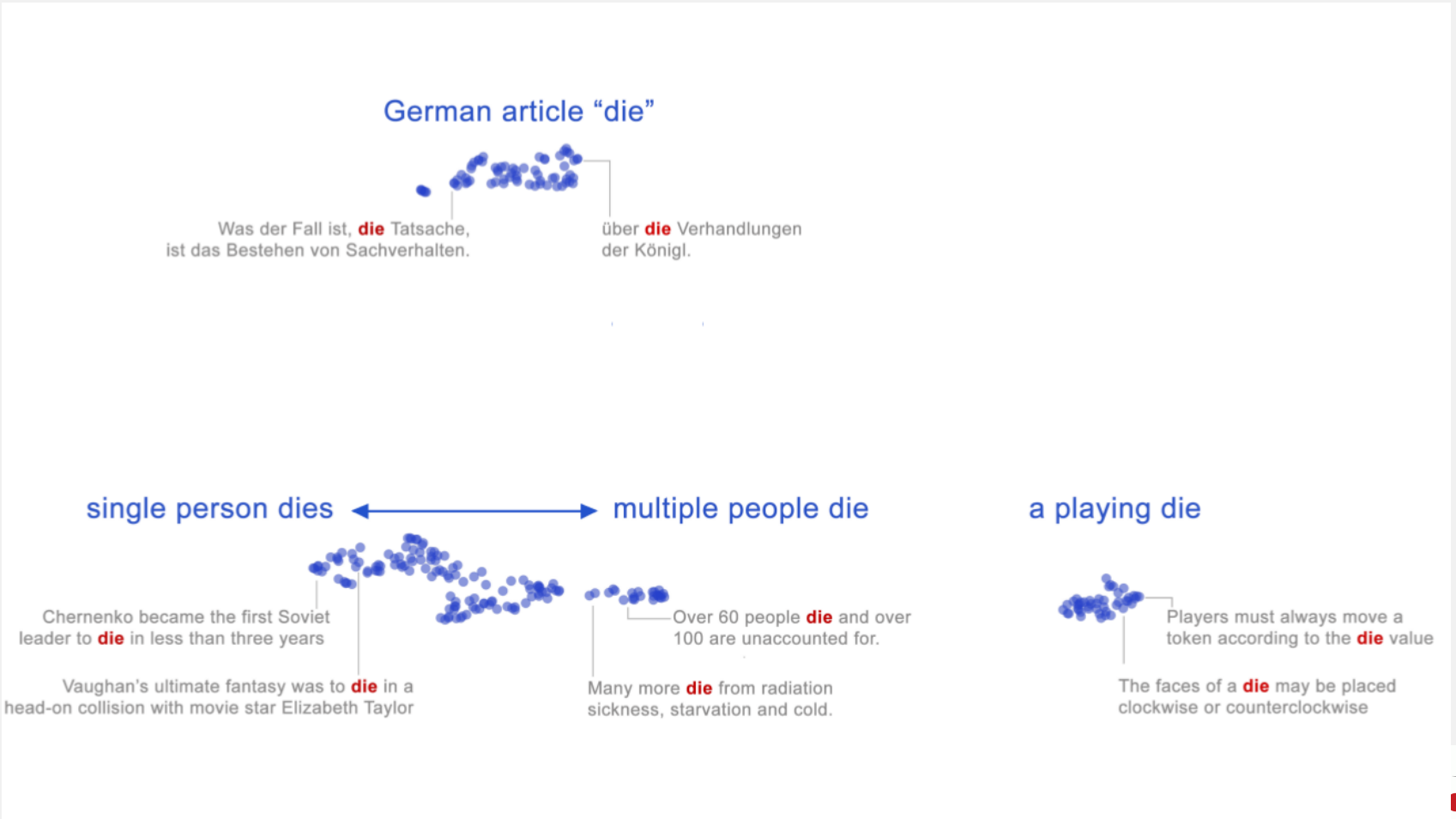
Homonymy and polysemy

- 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

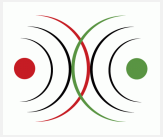
Polysemy in contextualized word embeddings

Coenen et al., (2019)



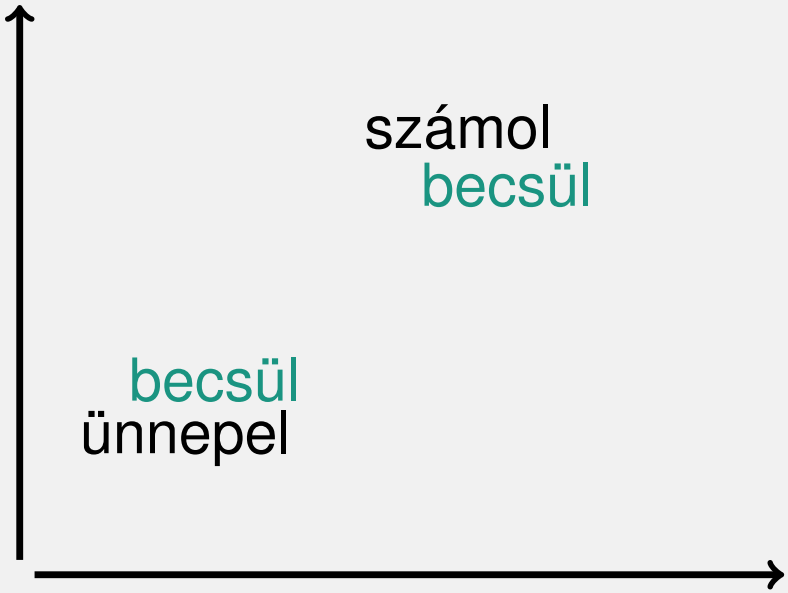
Multi-sense word embeddings (static)

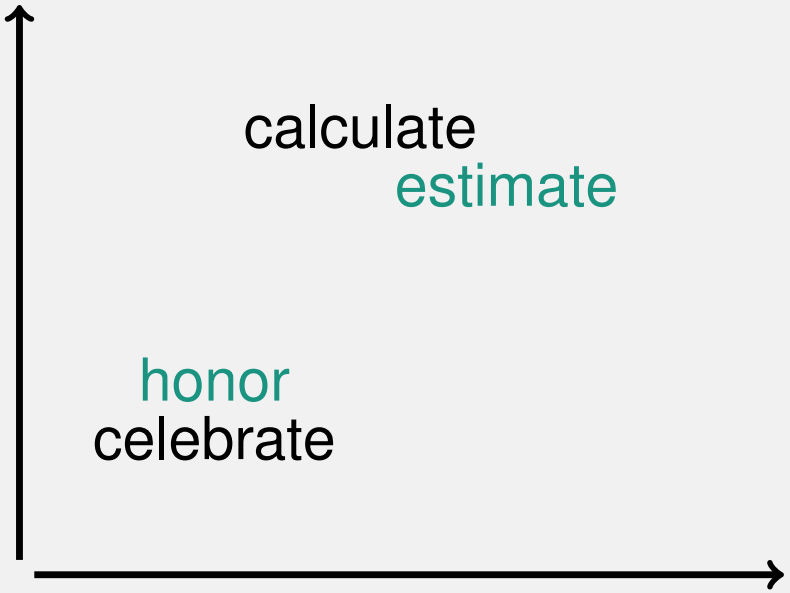
- before contextualized word representations, in a galaxy. . .
- multi-“prototype” embeddings (Reisinger and Mooney, 2010)
- with neural network (Huang et al., 2012)
- multi-sense skip gram, open-source tools
  - Neelakantan et al., 2014
  - as a Dirichlet Process
    - `AdaGram` (Bartunov et al., 2016)
    - `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 and noise vectors)



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)



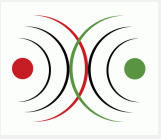


Data

- 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

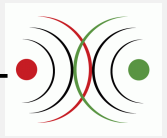


Overview

- Word embeddings
  - Static
  - Contextualized
- Word senses
  - Results

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



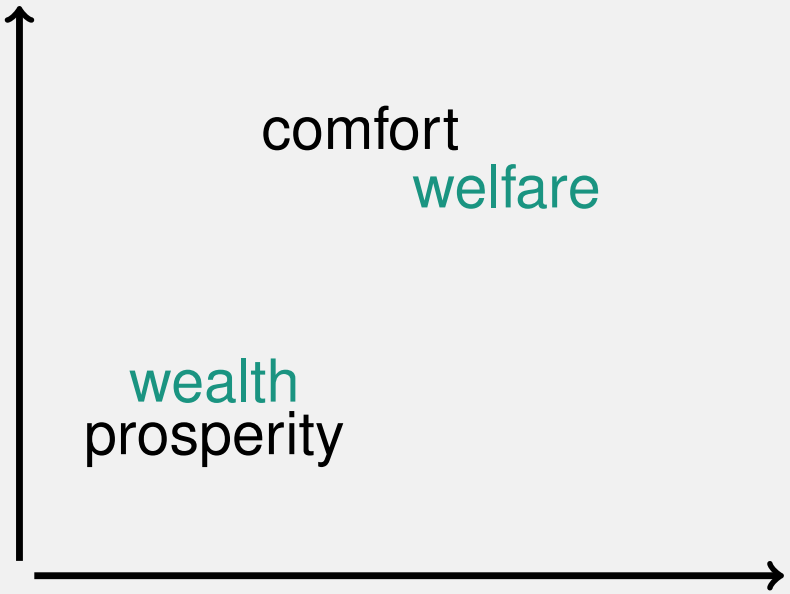
The resolution trade-off

- 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" (u_i)	71.0%	19.46%
`mutli` "context vectors" (v_i)	69.9%	**20.76%**

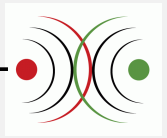
$$p(w_i \mid w_j) \propto \exp(u_i^\top v_j)$$

Problem: synonymous translations



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



Overview

Word embeddings
- Static
- Contextualized

Word senses
- Results

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. 18).


Bartunov, Sergey et al. (2016). “Breaking Sticks and Ambiguities with Adaptive Skip-gram”. In: *Proceedings of Machine Learning Research* 51: Artificial Intelligence and Statistics, pp. 130–138 (cit. on p. 16).


Belinkov, Yonatan and James Glass (2019). “Analysis methods in neural language processing: A survey”. In: *Transactions of the Association for Computational Linguistics* 7, pp. 49–72 (cit. on pp. 11, 12).


Bojanowski, Piotr, Armand Joulin, and Tomas Mikolov (2016). “Alternative Structures for Character-level RNNs”. In: *International Conference on Learning Representations, Workshop track (ICLR 2016)*. arXiv: `1511.06303 [cs.LG]` (cit. on p. 6).


Borbély, Gábor et al. (2016a). “Denoising composition in distributional semantics”. In: *DSALT: Distributional Semantics and Linguistic Theory*. poster (cit. on p. 18).

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. 17).
- Coenen, Andy et al. (2019). "Visualizing and Measuring the Geometry of BERT". In: *arXiv preprint arXiv:1906.02715* (cit. on pp. 11, 12, 15).
- Collobert, R. et al. (2011). "Natural Language Processing (Almost) from Scratch". In: *Journal of Machine Learning Research (JMLR)* (cit. on p. 6).
- Conneau, Alexis et al. (2018). "What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2126–2136. URL: http://aclweb.org/anthology/P18-1198 (cit. on p. 11).
- Şahin, Gözde Gül et al. (2019). *LINSPECTOR: Multilingual Probing Tasks for Word Representations*. arXiv:1903.09442 (cit. on p. 12).

Bibliography III

- Dalvi, Fahim et al. (2019). “What is one grain of sand in the desert? analyzing individual neurons in deep nlp models”. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (cit. on p. 11).
- Devlin, Jacob et al. (2018). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. Version 1. In: *arXiv preprint arXiv:1810.04805*. arXiv: 1810.04805v1 [cs.CL]. URL: http://arxiv.org/abs/1810.04805v1 (cit. on pp. 4, 10).
- Ethayarajh, Kawin (2019). “How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2”. In: *EMNLP* (cit. on p. 12).
- 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. 6).
- Hewitt, John and Christopher D Manning (2019). “A structural probe for finding syntax in word representations”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138 (cit. on p. 11).

Bibliography IV

- 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).
- 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. 16).
- Katz, J. and Jerry A. Fodor (1963). “The structure of a semantic theory”. In: *Language* 39, pp. 170–210 (cit. on p. 7).
- 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 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. 6).

Bibliography V

- Li, Jiwei and Dan Jurafsky (2015). "Do Multi-Sense Embeddings Improve Natural Language Understanding?" In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 1722–1732. DOI: 10.18653/v1/D15-1200. URL: http://www.aclweb.org/anthology/D15-1200 (cit. on p. 16).
- Mikolov, Tomas, Quoc V Le, and Ilya Sutskever (2013). "Exploiting similarities among languages for machine translation". arXiv preprint arXiv:1309.4168 (cit. on p. 8).
- 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. 7).
- 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: https://bit.ly/39HikH8 (cit. on p. 6).

Bibliography VI

- Mikolov, Tomas et al. (2013b). “Efficient Estimation of Word Representations in Vector Space”. In: *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*. Ed. by Y. Bengio and Y. LeCun. arXiv: 1301.3781 [cs.CL]. URL: http://arxiv.org/abs/1301.3781 (cit. on p. 17).
- Neelakantan, Arvind et al. (2014). “Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1059–1069. DOI: 10.3115/v1/D14-1113. URL: http://www.aclweb.org/anthology/D14-1113 (cit. on p. 16).
- Nemeskey, Dávid Márk (2017). “emLam – a Hungarian Language Modeling baseline”. In: *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*. Szeged, pp. 91–102. arXiv: 1701.07880 [cs.CL] (cit. on p. 18).
- Oravecz, Csaba, Tamás Váradi, and Bálint Sass (2014). “The Hungarian Gigaword Corpus”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavik, Iceland: European Language Resources Association (ELRA). URL: http://www.aclweb.org/anthology/L14-1536 (cit. on p. 18).

Bibliography VII


Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). “Glove: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: http://www.aclweb.org/anthology/D14-1162 (cit. on pp. 17, 18).


Peters, Matthew et al. (2018a). “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: http://aclweb.org/anthology/N18-1202 (cit. on pp. 4, 6, 10).


Peters, Matthew et al. (2018b). “Dissecting Contextual Word Embeddings: Architecture and Representation”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1499–1509. URL: http://aclweb.org/anthology/D18-1179 (cit. on p. 11).

Bibliography VIII


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. 16).


Schütze, Hinrich (1998). “Automatic Word Sense Discrimination”. In: *Computational Linguistics Special-Issue-on-Word Sense Disambiguation* 24.1. URL: http://www.aclweb.org/anthology/J98-1004 (cit. on p. 14).


Vaswani, Ashish et al. (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 5998–6008. arXiv: 1706.03762 [cs.CL]. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf (cit. on p. 6).


Wang, Bin and C.-C. Jay Kuo (2020). *SBERT-WK: A Sentence Embedding Method by Dissecting BERT-based Word Models*. arXiv: 2002.06652 [cs.CL].