The Third Workshop on Evaluating Vector Space Representations for NLP

Anna Rogers, Aleksandr Drozd, Anna Rumshisky, Yoav Goldberg

June 6, 2019

Co-located with NAACL 2019, Minneapolis, USA

RepEval 2016 @ACL

- Organizers
  - Omer Levy
  - Felix Hill
  - Anna Korhonen
  - Kyunghyun Cho
  - Roi Reichart
  - Yoav Goldberg
  - Antoine Bordes
- analysis track + proposal track
- 39 submissions, 16 accepted (5 in the analysis track, 41% acceptance)
- $\approx$ 150 attendees

RepEval 2017 @EMNLP

- Organizers
  - Sam Bowman
  - Yoav Goldberg
  - Felix Hill
  - Angeliki Lazaridou
  - Omer Levy
  - Roi Reichart
  - Anders Søgaard
- proposal track, MultiNLI shared task (to evolve into GLUE)
- 16 submissions, 11 accepted (68.8% acceptance)
- $\approx$ 250 attendees

Do we even need word embeddings anymore?

RepEval 2019 @ NAACL

- Organizers
  - Anna Rogers
  - Aleksandr Drozd
  - Anna Rumshisky
  - Yoav Goldberg
- analysis track + proposal track
- 25 submissions (+ 2 withdrawn), 13 accepted (52% acceptance)

RepEval 2019: Program Committee

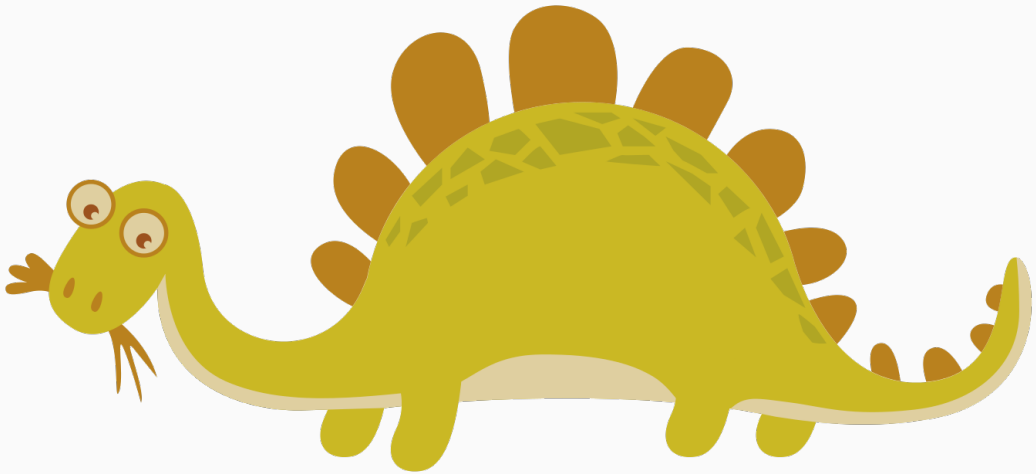
- Omri Abend
- Emily Bender
- Sam Bowman
- Jose Camacho Collados
- Alexis Conneau
- Barry Devereux
- Georgiana Dinu
- Allyson Ettinger
- Mohit Iyyer
- Hila Gonen
- Douwe Kiela
- Jonathan K. Kummerfeld
- Tal Linzen
- Preslav Nakov
- Neha Nayak
- Mark Neumann
- Ellie Pavlick
- Denis Paperno
- Marek Rei
- Roi Reichart
- Vered Shwartz
- Diarmuid O'Seaghdha
- Gabriel Stanovsky
- Karl Stratos
- Yulia Tsvetkov
- Ivan Vulić
- Luke Zettlemoyer

Vector Meaning Representations: 6 Years Later

A brief and biased overview

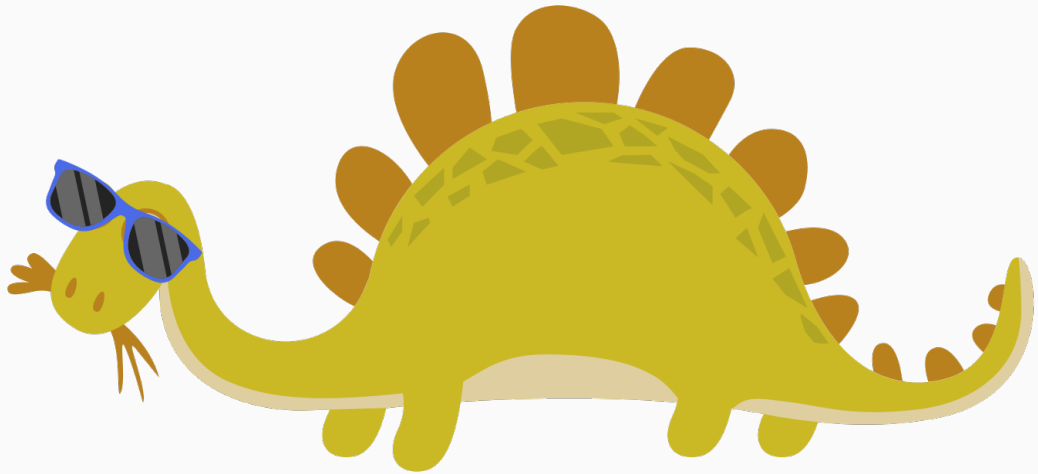
When the earth was still flat...

- distributional hypothesis (Firth, 1957; Harris, 1954) $\Rightarrow$ corpus linguistics work on word association measures;
- count-based distributional meaning representations, sparse and with reduced dimensionality with PCA, PPMI, SVD...
- sem. spaces in psycholinguistics: LSA (Landauer et al., 1998), HAL (Lund and Burgess, 1996), ICA (Väyrynen and Honkela, 2004)...
- work on DSM compositionality (Mitchell and Lapata, 2008, 2010; Baroni and Zamparelli, 2010; Baroni, 2013; Lazaridou et al., 2013)

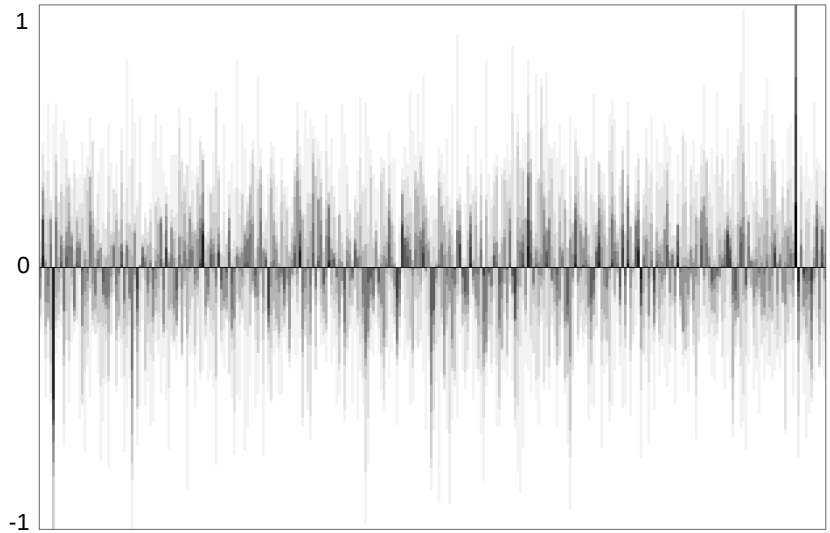


And then deep learning came

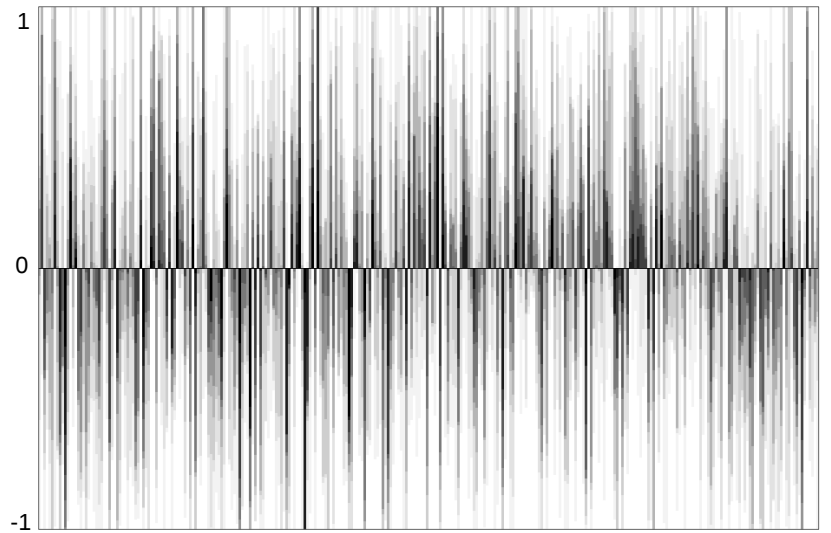
- word2vec (Mikolov et al., 2013a,b)
- Don't count, predict! (Baroni et al., 2014)
- GloVe (Pennington et al., 2014)



Something meaningful is going on!



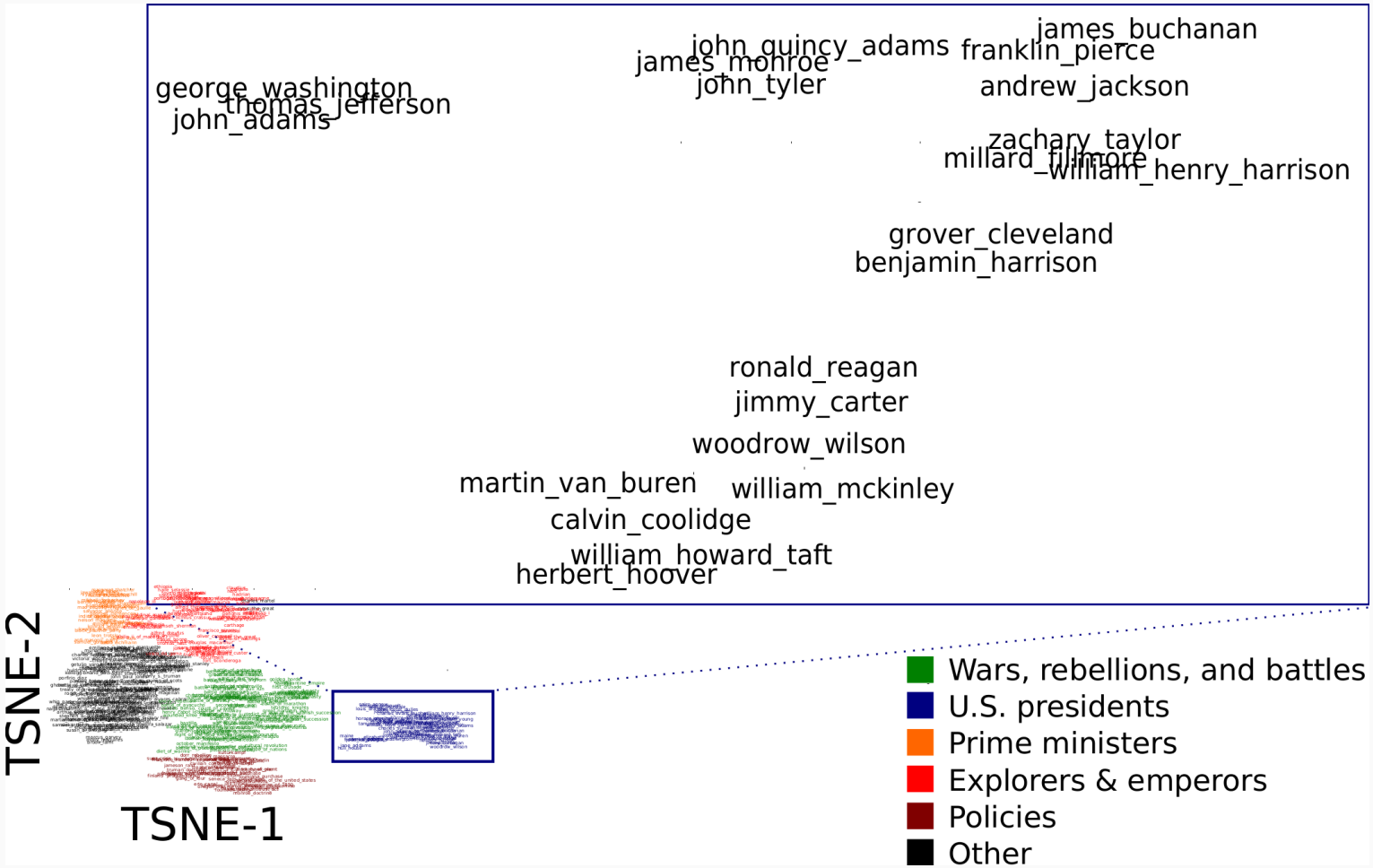
10 random words: *emergency, bluff, buffet, horn, human, like, american, pretend, tongue, green*



10 felines: *cat, lion, tiger, leopard, cougar, cheetah, lynx, bobcat, panther, puma*

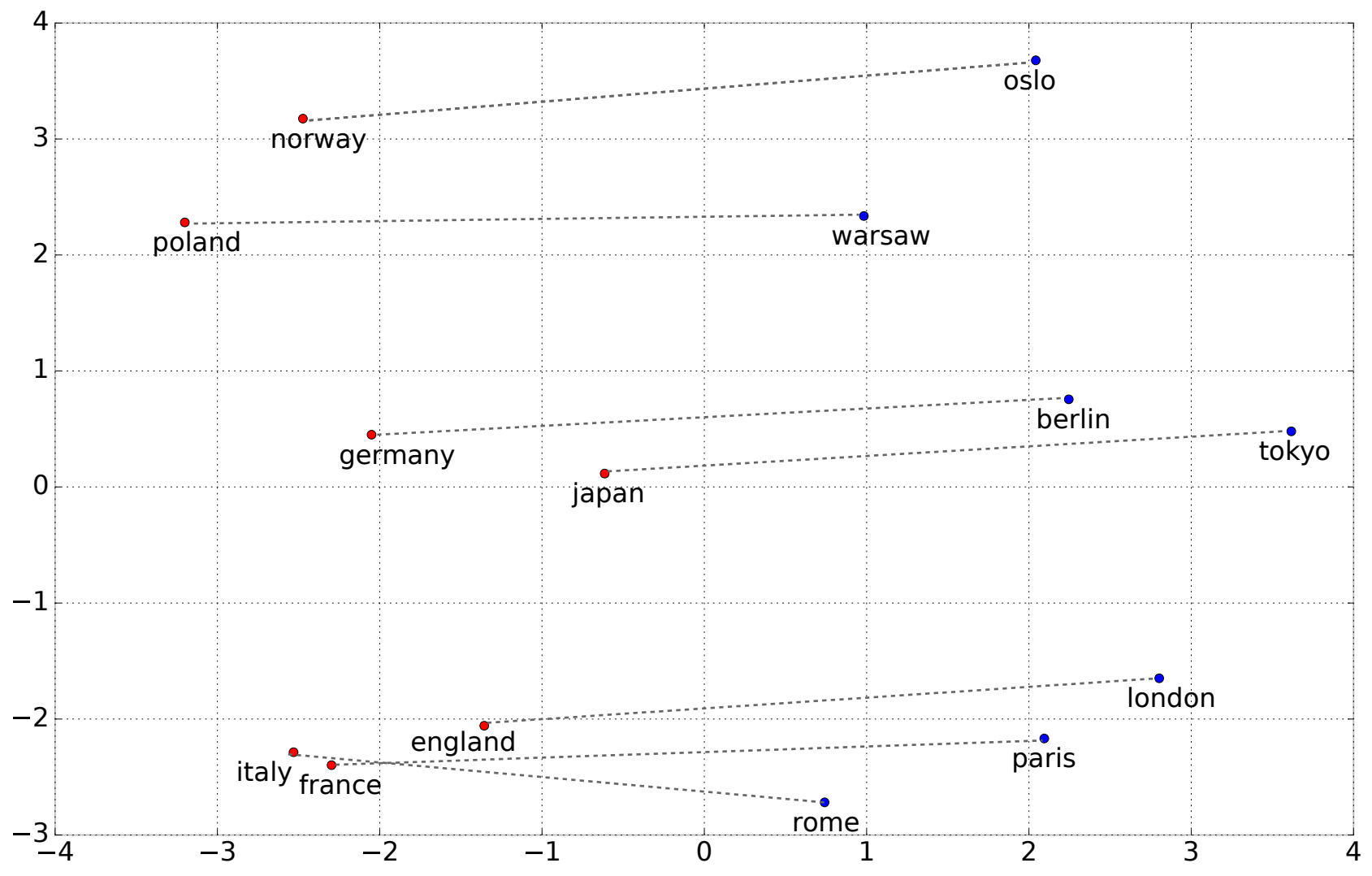
GloVe visualization (Gladkova and Drozd, 2016)

Something meaningful is going on!



Iyyer et al. (2014)

Have we *solved* meaning?



GloVe (Pennington et al., 2014)

Let's extend that!

- subword embeddings (Bojanowski et al., 2017; Cotterell and Schütze, 2015):
- subcharacter embeddings (Sun et al., 2014; Yu et al., 2017; Stratos, 2017; Karpinska et al., 2018):
- syntax-aware embeddings (Levy and Goldberg, 2014a; Li et al., 2017; Lapesa and Evert, 2017):
- retrofitted embeddings (Faruqui et al., 2016; Mrkšić et al., 2016; Yu et al., 2016)
- sentence embeddings (Kiros et al., 2015; Conneau et al., 2017; Bowman et al., 2016; Hill et al., 2016; Le and Mikolov, 2014)

The black box is not entirely magic

- Levy and Goldberg (2014b): Neural word embedding as implicit matrix factorization
- Lebret and Collobert (2015): you're just not using PCA right!
- Overall similar behavior with SVD on analogy task (Gladkova et al., 2016)

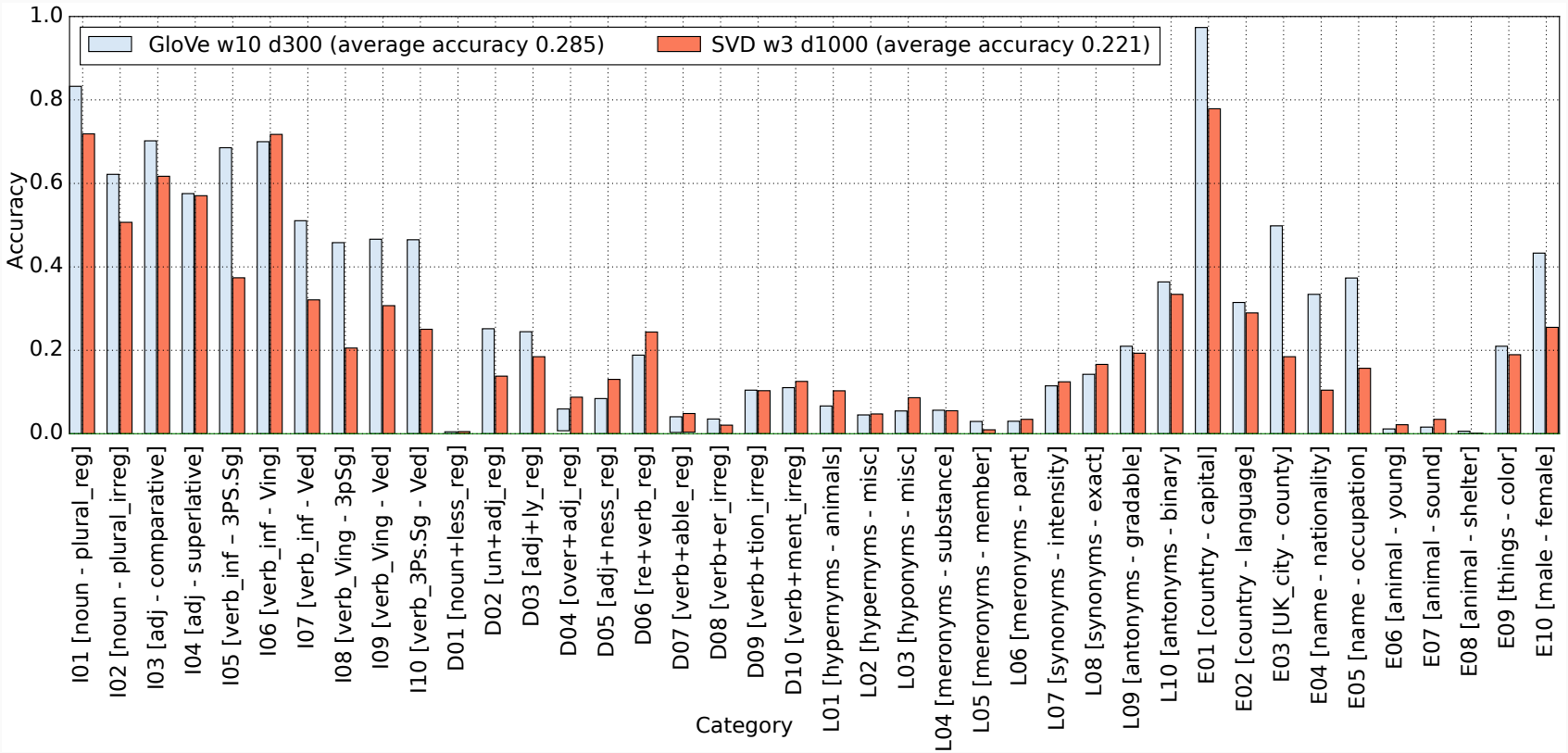
Relatedness/similarity is not a great metric

WordSim353
tiger cat 7.35
book paper 7.46
computer keyboard 7.62
plane car 5.77
train car 6.31
telephone communication 7.50
television radio 6.77
media radio 7.42
drug abuse 6.85
cucumber potato 5.92
bread butter 6.19
doctor nurse 7.00
smart student 4.62
smart stupid 5.81

- task with a long history (Geffet and Dagan, 2004; Turney, 2006; Agirre et al., 2009; Kotlerman et al., 2010)
- WordSim353 (Finkelstein, Garilovich et al. 2002), MEN (Bruni, Tran, and Baroni, 2013), RareWords (Luong, Socher and Manning, 2013), Radinsky Mturk (Radinsky, Agichtein et al., 2011))
- relatedness vs similarity (Hill et al., 2015b; Kiela et al., 2015)
- Methodological problems (Gladkova and Drozd, 2016; Faruqui et al., 2016), x10 for text

No, we don't really have analogical reasoning

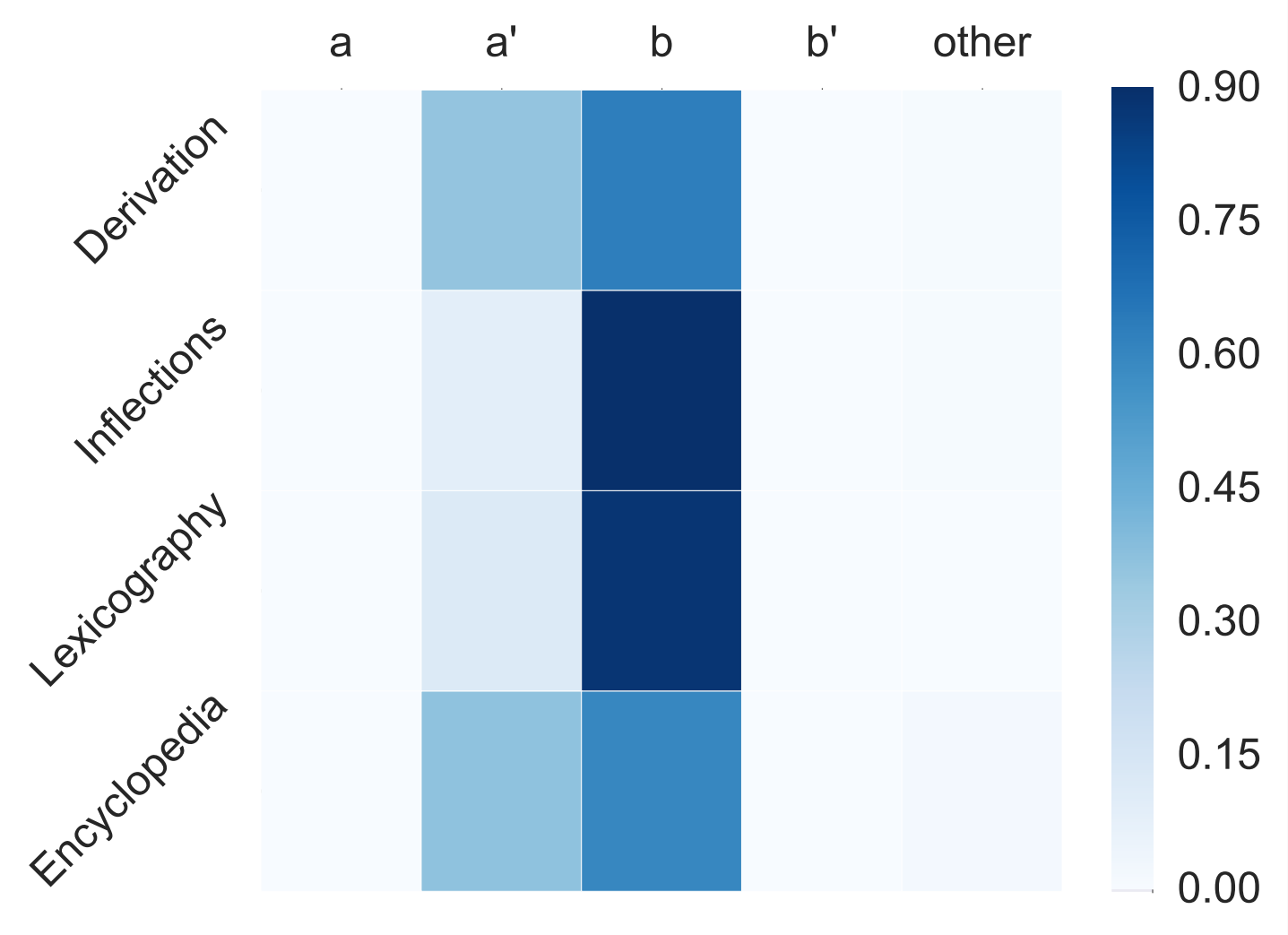
$\overrightarrow{Berlin}$ - $\overrightarrow{Germany}$ + $\overrightarrow{Japan}$ = $\overrightarrow{Tokyo}$ (Mikolov et al. 2013)



Bigger Analogy Test Set (Gladkova et al., 2016)

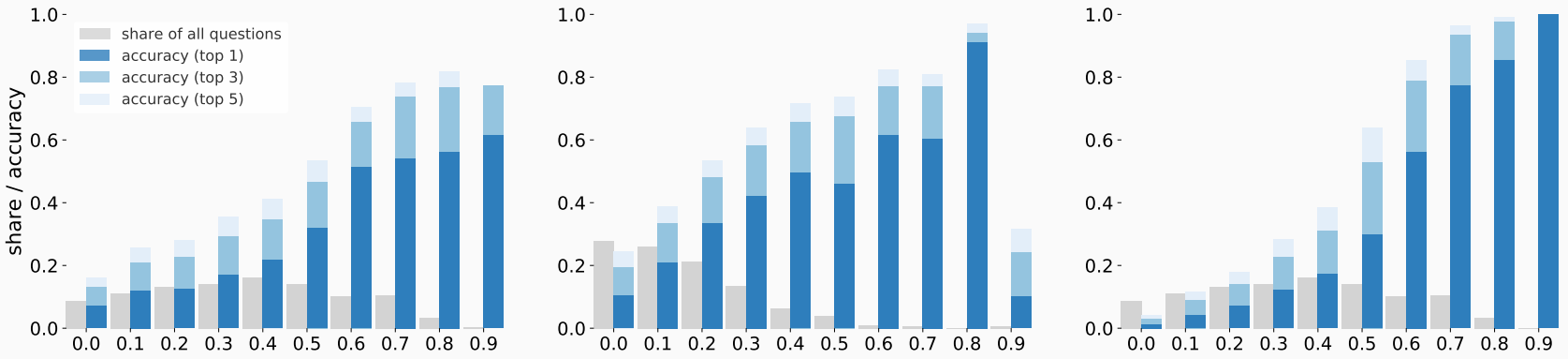
Similar results for Japanese (Karpinska et al., 2018)

Mikolov cheated! (Rogers et al., 2017)



The "honest" solution to $a' - a + b$

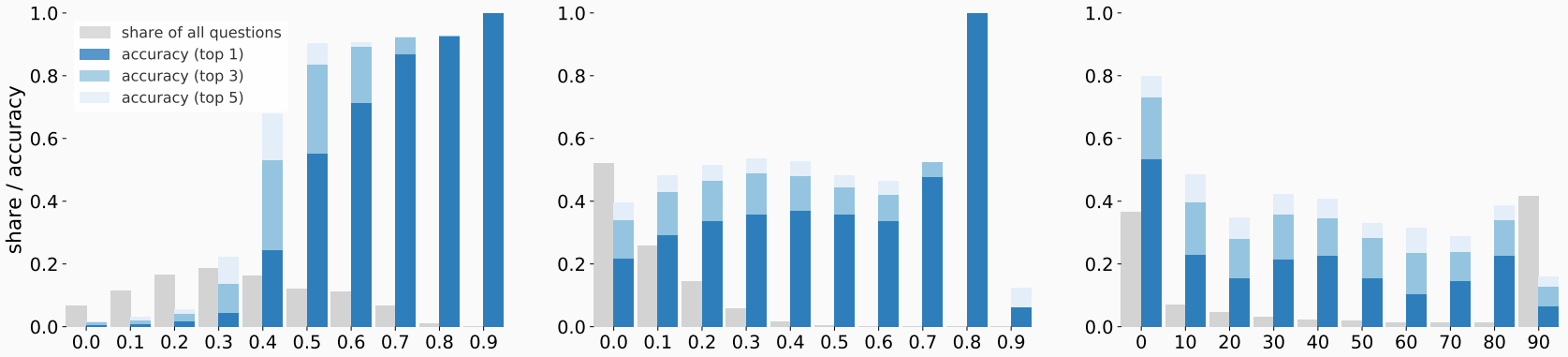
Cosine similarity bias in word analogies (Rogers et al., 2017)



(a) similarity between vectors a and a'

(b) similarity between vectors a' and b'

(c) similarity between vectors b and b'



(d) similarity between vector b' and predicted vector

(e) similarity between vector b' and a

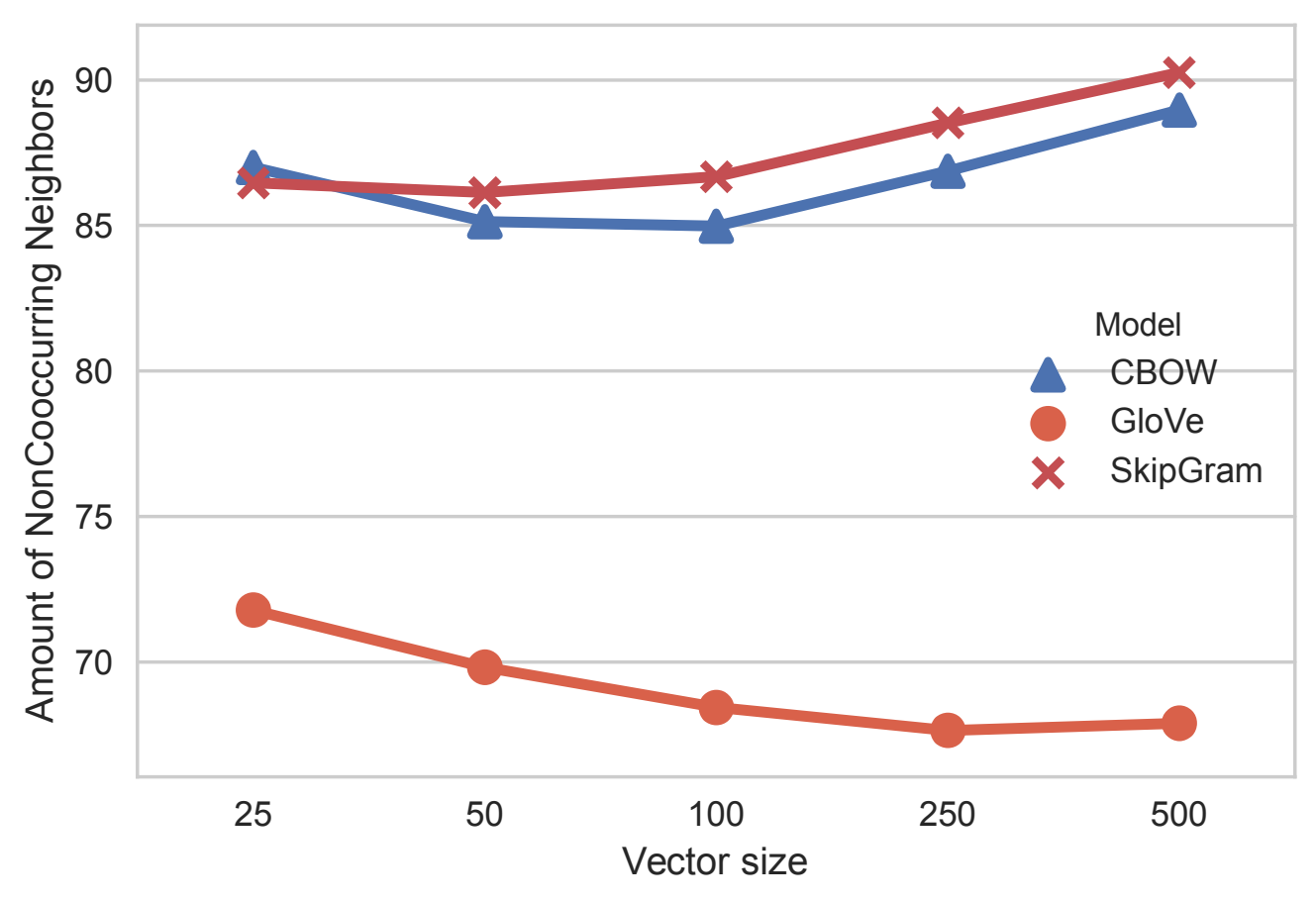
(f) rank of b in the neighborhood of b'

Vector offset method accuracy by cosine similarity bins (GloVe)

Parameters matter a LOT

- Levy et al. (2015): parameters can matter more than the model
- let's study parameters! (Lapesa and Evert, 2014; Lai et al., 2016; Wielfaert et al., 2014; Kiela and Clark, 2014; Melamud et al., 2016b)

Parameters (Rogers et al., 2018)



Detection of word relations without corpus evidence: vector size effect

The shift to extrinsic evaluations

Intrinsic evaluations fail to predict task performance (Chiu et al., 2016; Rogers et al., 2018) $\Rightarrow$

1. “Representative suite of extrinsic tasks" (Nayak et al., 2016)
2. SentEval (Conneau and Kiela, 2018) (partly);
3. GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019);

Quest for high-level reasoning: explosion of QA datasets

- *open-domain QA*: Natural Questions (Kwiatkowski et al., 2019), SearchQA (Dunn et al., 2017), MS MARCO (Nguyen et al.), TriviaQA (Joshi et al., 2017).

- *Extractive RC datasets*: SQuAD (Rajpurkar et al., 2016, 2018), WikiQA (Yang et al., 2015), WikiLinks Rare Entity Prediction (Long et al., 2017), CBT (Hill et al., 2015a), BookTest (Bajgar et al., 2017), MCTest (Richardson et al., 2013), NewsQA (Trischler et al., 2016), CNN/Daily Mail (Hermann et al., 2015), Who Did What (Onishi et al., 2016).

- *Academic QA tests*: RACE (Lai et al., 2017), OpenBookQA (Mihaylov et al., 2018), CLEF QA (Peñas et al., 2014), ARC (Clark et al., 2018);

- *QA involving commonsense knowledge*: MCScript (Ostermann et al., 2018), RocStories (Mostafazadeh et al., 2017), CommonsenseQA (Talmor et al., 2019);

- QA with reasoning over over *long texts* (Kocisky et al., 2018) and *multiple documents*: HotpotQA (Yang et al., 2018), QAngaroo (Welbl et al., 2018), ComplexWebQuestions (Talmor and Berant, 2018);

- Other: QuAC (Choi et al., 2018), CoQA (Reddy et al., 2018), BoolQ (Clark et al., 2019), DROP (Dua et al., 2019) ...

Well, not so high-level, actually

- human-level performance on SQuAD can be achieved while relying only on superficial cues (Jia and Liang, 2017);
- 73% of the NewsQA can be solved by simply identifying the single most relevant sentence (Chen et al., 2016);
- in the commonsense reasoning challenge of SemEval2018-Task 11 (Ostermann et al., 2018) most participants did not use any extra knowledge sources, and one of them still achieved 0.82 accuracy vs 0.84 achieved by the winner;
- models trained on one dataset do not necessarily do well on another, even in the same domain (Yatskar, 2019).

.

Ditto for inference

- *NLI datasets*: SNLI (Williams et al., 2017), MultiNLI(Nangia et al., 2017), DialogueNLI (Welleck et al., 2018), MedNLI (Romanov and Shivade, 2018), SciTail (Khot et al.), JHU Ordinal Common-sense Inference (Zhang et al., 2017), SWAG (Zellers et al., 2018) (+ all the RTE datasets)
- *problems with NLI*: Glockner et al. (2018); Gururangan et al. (2018); Poliak et al. (2018); McCoy et al. (2019)

Are we scoring high/low due to representation or method?

Solving BATS word analogies: accuracy for 3 methods
(Drozd et al., 2016)

Method	Encyclopedia		Lexicography		Inflections		Derivation	
	GloVe	SG	GloVe	SG	GloVe	SG	GloVe	SG
3CosAdd	31.5%	26.5%	10.9%	9.1%	59.9%	61.0%	10.2%	11.2%
3CosAvg	**44.8%**	34.6%	13.0%	9.6%	68.8%	69.8	11.2%	15.2%
LRCos	40.6%	**43.6%**	16.8%	**15.4%**	74.6%	87.2%	17.0%	**45.6%**

If we have credit problem with analogies, what about high-level tasks?

Are we scoring high/low due to representation or method?

Solving BATS word analogies: accuracy for 3 methods
(Drozd et al., 2016)

Method	Encyclopedia		Lexicography		Inflections		Derivation	
	GloVe	SG	GloVe	SG	GloVe	SG	GloVe	SG
3CosAdd	31.5%	26.5%	10.9%	9.1%	59.9%	61.0%	10.2%	11.2%
3CosAvg	**44.8%**	34.6%	13.0%	9.6%	68.8%	69.8	11.2%	15.2%
LRCos	40.6%	**43.6%**	16.8%	**15.4%**	74.6%	87.2%	17.0%	**45.6%**

a representation with information X readily available

$\Rightarrow$

better performance on task Y

Linguistic diagnostics methodology: what kind of information does your representation prioritize?

1. Choose vocabulary important for your task, or a general representative sample.

in this study: 908 verbs, nouns, adjectives, adverbs, balanced by POS and frequency

2. Get top n neighbors and similarity scores.

Rank	Deps		FastText	
1	colour	0.93	$color	0.75
2	colors	0.72	color...	0.69
3	coloration	0.69	colour	0.69
4	colouration	0.68	color#ff	0.69
5	colours	0.68	color#d	0.68
6	hue	0.66	@color	0.67
7	hues	0.65	barcolor	0.67

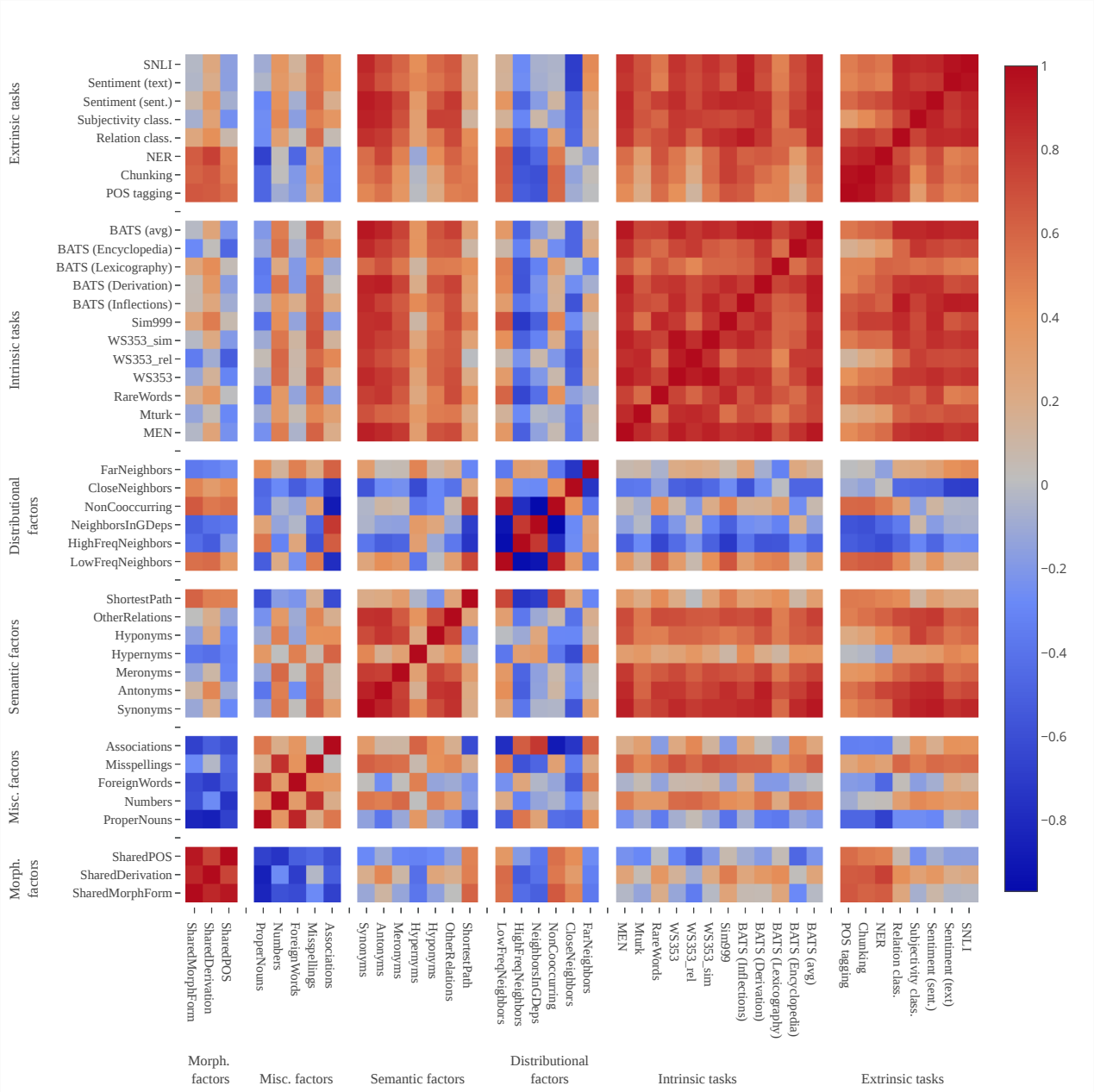
3. Annotate linguistic relations in the vector neighborhoods.

color: colors -> inflected form
color: hue -> synonym
color: coloration -> derived term

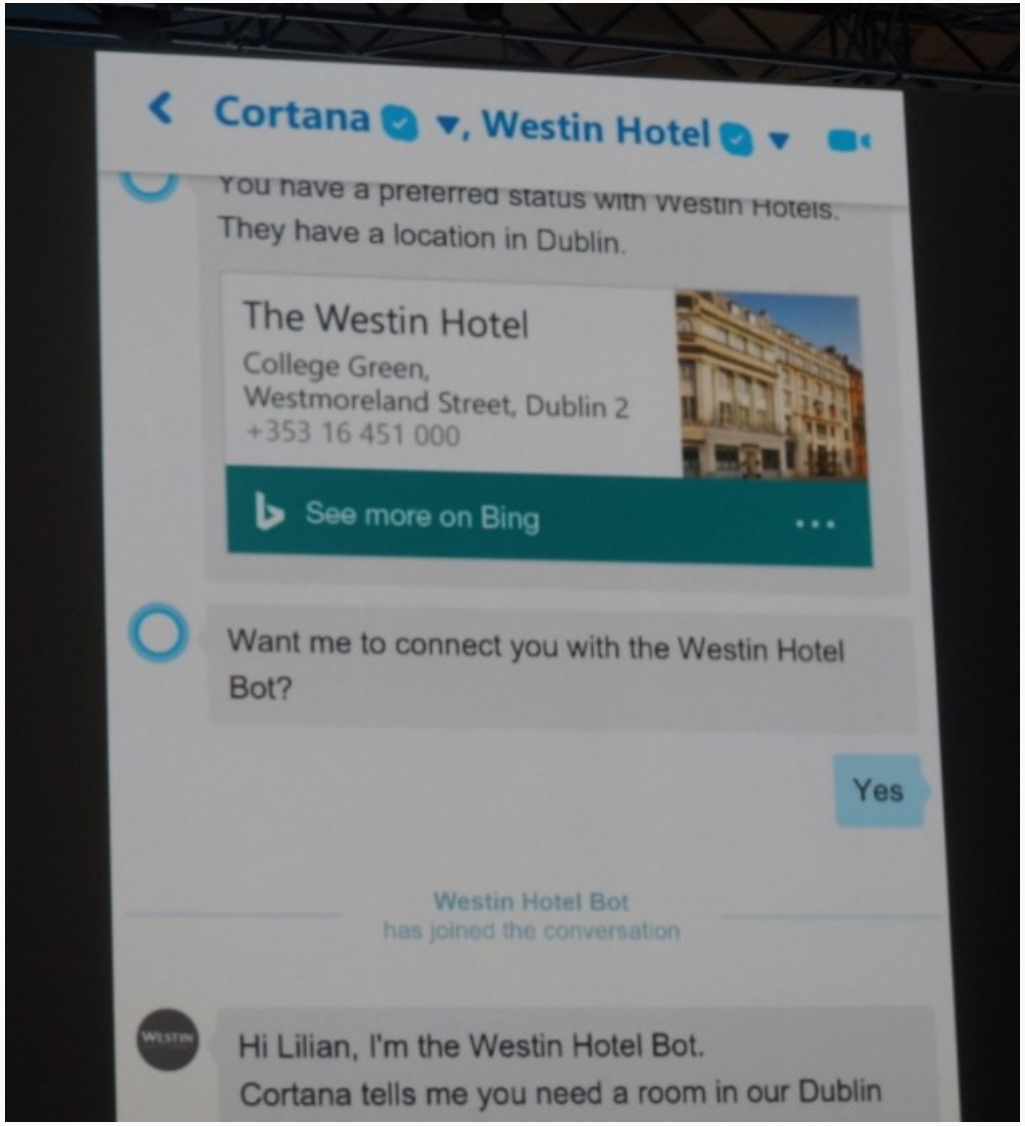
4. Compare models/parameters, adjust, repeat.

FastText: X synonyms, Y antonyms
DEPS: Y synonyms, Z antonyms

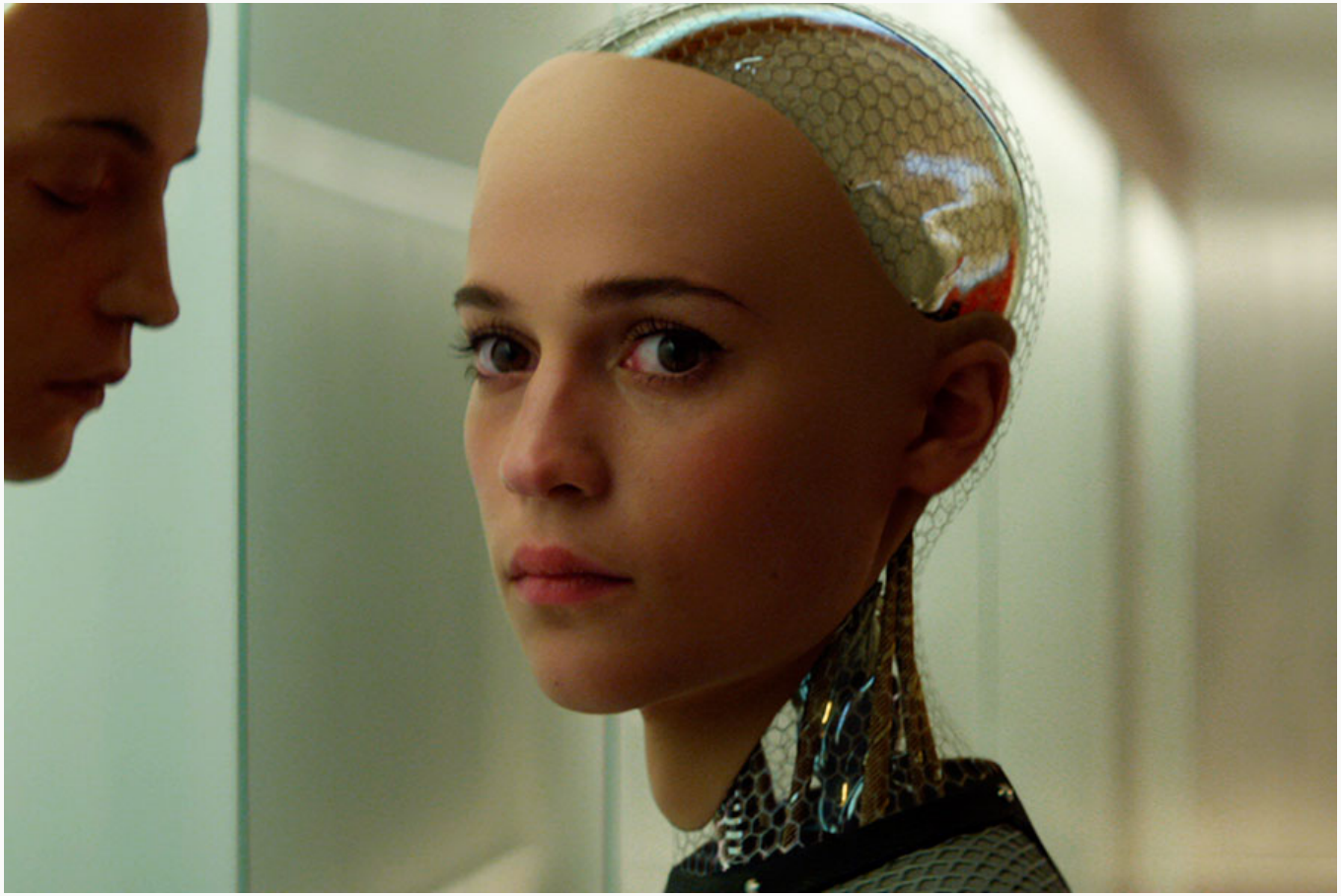
No free lunch: specialized neighbors -> performance (Rogers et al., 2018)



Specialization is great for industrial applications...



... but it won't get us to general AI



Reproducibility crisis

- instability in learned word embeddings (Wendlandt et al., 2018; Antoniak and Mimno, 2018; Pierrejean and Tanguy, 2018);
- variability of results by deep learning methods (Crane, 2018);
- misattribution of impact due to pipeline components;

All of that in a field fighting for +2% gain over SOTA

Push for interpretability

- interpretable dimensions (Nalisnick and Ravi, 2015; Sun et al., 2016; Fyshe et al., 2015)
- linguistically-motivated evaluation of meaning representations (Tsvetkov et al., 2016; Rogers et al., 2018);
- probing for linguistic structures (Ettinger et al., 2016; Liu et al., 2019b; Conneau and Kiela, 2018; Wang et al., 2018; Strubell and McCallum, 2018)
- Workshops: Relevance of Linguistic Structure in Neural NLP (ACL 2018), Workshop on Evaluating Vector Space Representations for NLP (ACL 2016, EMNLP20 17, NAACL 2019), Building Linguistically Generalizable NLP Systems (EMNLP 2017), Workshop on Designing Meaning Representations (ACL 2019), Blackbox NLP (ACL 2019)

Who wants word embeddings anymore?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 Apr 13, 2019	SemBERT(ensemble) *Shanghai Jiao Tong University*	86.166	88.886
4 May 14, 2019	SG-Net (ensemble) *Anonymous*	86.211	88.848
5 Mar 16, 2019	BERT + DAE + AoA (single model) *Joint Laboratory of HIT and iFLYTEK Research*	85.884	88.621
6 May 14, 2019	SG-Net (single model) *Anonymous*	85.229	87.926
7 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	85.150	87.715

Who wants word embeddings anymore?

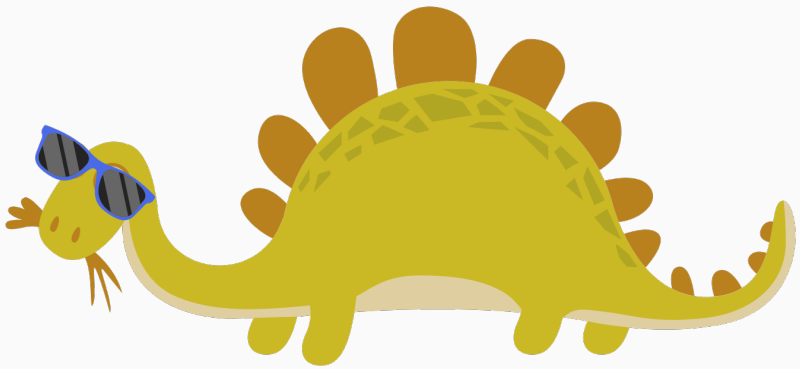
- sense-aware extensions of word2vec (Neelakantan et al., 2014; Liu et al., 2015; Piña and Johansson, 2015; Lee and Chen, 2017)
- early models combining sense and context representaitons (Li and McCallum, 2005; Melamud et al., 2016a)
- TagLM (Peters et al., 2017), CoVe (McCann et al., 2017), ELMO (Bowman et al., 2018), BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019)

Current problems of contextualized representations

- likely overparametrization (Frankle and Carbin, 2018; Goldberg, 2019; Adhikari et al., 2019; Wu et al., 2019)
- interpretability (Goldberg, 2019; Jawahar et al., 2019; Tran et al., 2018; Liu et al., 2019a)
- too computationally demanding for people in academia to experiment a lot with (and to keep up with the industry)
- scaring away people from other disciplines

Thank You!

Slides: `http://cs.uml.edu/~arogers/` ("Talks" tab)



References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Rethinking complex neural network architectures for document classification. In *Proceedings of NAACL 2019: Conference of the North American Chapter of the Association for Computational Linguistics*.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pașca, and Aitor Soroa. 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association of Computational Linguistics*, 6:107–119.

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2017. Embracing data abundance: BookTest Dataset for Reading Comprehension. In *ICLR*.

Marco Baroni. 2013. Composition in Distributional Semantics. *Language and Linguistics Compass*, 7(10):511–522.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting

semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, MIT, Massachusetts, USA, 9-11 October 2010.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5(0):135–146.

Samuel R. Bowman, Ellie Pavlick, Edouard Grave, Benjamin Van Durme, Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, and Berlin Chen. 2018. Looking for ELMo's Friends: Sentence-Level Pretraining Beyond Language Modeling. *arXiv:1812.10860 [cs]*.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating Sentences from a Continuous Space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural*

Language Learning (CoNLL), pages 10–21, Berlin, Germany, August 7-12, 2016. Association for Computational Linguistics.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367. Association for Computational Linguistics.

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance. pages 1–6. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. 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), pages 2924–2936.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457 [cs]*.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 7–11, 2017. Association for Computational Linguistics.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological Word-Embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, pages 1287–1292.

Matt Crane. 2018. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. *Transactions of the Association for Computational Linguistics*, 6:241–252.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka, Japan, December 11-17.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. 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)*, pages 2368–2378.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. *arXiv:1704.05179 [cs]*.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35.

J. R. Firth. 1957. A synopsis of linguistic theory 1930-55. 1952-59:1–32.

Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*.

Alona Fyshe, Leila Wehbe, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. 2015. A Compositional and Interpretable Semantic Space. *Proceedings of the NAACL-HLT, Denver, USA*.

Maayan Geffet and Ido Dagan. 2004. Feature Vector Quality and Distributional Similarity. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of The 1st Workshop on Evaluating Vector Space Representations for NLP*, pages 36–42, Berlin, Germany. ACL.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL-HLT SRW*, pages 47–54, San Diego, California, June 12-17, 2016. ACL.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA. MIT Press.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015a. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *arXiv:1511.02301 [cs]*.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning Distributed Representations of Sentences from Unlabelled Data. In *Proceedings of NAACL-HLT 2016*, pages 1367–1377, San Diego, California, June 12-17, 2016. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015b. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. 2014. A Neural Network for Factoid Question Answering over Paragraphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 633–644, Doha, Qatar. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy.

Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611. Association for Computational Linguistics.

Marzena Karpinska, Bofang Li, Anna Rogers, and Aleksandr Drozd. 2018. Subcharacter Information in Japanese Embeddings: When Is It Worth It? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 28–37, Melbourne, Australia. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. SCITAIL: A Textual Entailment Dataset from Science Question Answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5189–5197.

Douwe Kiela and Stephen Clark. 2014. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and Their Compositionality (CVSC) at EACL*, pages 21–30.

Douwe Kiela, Felix Hill, and Stephen Clark. 2015. Specializing Word Embeddings for Similarity or Relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*

(EMNLP)), pages 2044–2048, Lisbon, Portugal, 17-21 September 2015. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, volume 2, pages 3294–3302, Montreal, Canada, December 07 - 12, 2015.

Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794. Association for Computational Linguistics.

Siwei Lai, Kang Liu, Shizhu He, and Jun Zhao. 2016. How to generate a good word embedding. *IEEE Intelligent Systems*, 31(6):5–14.

Thomas K Landauer, Peter W Folt, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2):259–284.

Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545.

Gabriella Lapesa and Stefan Evert. 2017. Large-scale evaluation of dependency-based DSMs: Are they worth the effort? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 394–400. Association for Computational Linguistics.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics. In *ACL (1)*, pages 1517–1526.

Qv Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *International Conference on Machine Learning - ICML 2014*, volume 32, pages 1188–1196.

Rémi Lebret and Ronan Collobert. 2015. Rehabilitation of Count-based Models for Word Vector Representations. In *Computational Linguistics and Intelligent Text Processing*, pages 417–429. Springer.

Guang-He Lee and Yun-Nung Chen. 2017. MUSE: Modularizing Unsupervised Sense Embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 327–337.

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Bofang Li, Tao Liu, Zhe Zhao, Buzhou Tang, Aleksandr Drozd, Anna Rogers, and Xiaoyong Du. 2017. Investigating different syntactic context types and context representations for learning word embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2411–2421, Copenhagen, Denmark, September 7–11, 2017.

Wei Li and Andrew McCallum. 2005. Semi-supervised Sequence Modeling with Syntactic Topic Models. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*, AAAI'05, pages 813–818. AAAI Press.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A Smith. 2019a. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019b. Linguistic Knowledge and Transferability of Contextual Representations. In *NAACL*. Association for Computational Linguistics.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2015. Learning Context-sensitive Word Embeddings with Neural Tensor Skip-gram Model. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1284–1290. AAAI Press.

Teng Long, Emmanuel Bengio, Ryan Lowe, Jackie Chi Kit Cheung, and Doina Precup. 2017. World Knowledge for Reading Comprehension: Rare Entity Prediction with Hierarchical LSTMs Using External Descriptions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 825–834. Association for Computational Linguistics.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in Translation: Contextualized Word Vectors. *arXiv:1708.00107 [cs]*.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *arXiv:1902.01007 [cs]*.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016a. Context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61.

Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016b. The Role of Context Types and Dimensionality in Learning Word Embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1030–1040.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic

Regularities in Continuous Space Word Representations. In *Proceedings of NAACL-HLT 2013*, pages 746–751, Atlanta, Georgia, 9–14 June 2013.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based Models of Semantic Composition. In *ACL*, pages 236–244.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Nasrin Mostafazadeh, Michael Roth, Nathanael Chambers, and Annie Louis. 2017. LSDSem 2017 Shared Task: The Story Cloze Test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-Level Semantics*, pages 46–51. Association for Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting Word Vectors to Linguistic Constraints. In *Proceedings of NAACL-HLT 2016*, pages 142–148. Association for Computational Linguistics.

Eric Nalisnick and Sachin Ravi. 2015. Learning the Dimensionality of Word Embeddings. *arXiv:1511.05392 [cs, stat]*.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel R. Bowman. 2017. The Repeval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of the 2nd Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–10, Copenhagen, Denmark, September 7–11, 2017. Association for Computational Linguistics.

Neha Nayak, Gabor Angeli, and Christopher D. Manning. 2016. Evaluating Word Embeddings Using a Representative Suite of Practical Tasks. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, pages 19–23, Berlin, Germany, August 12, 2016. Association for Computational Linguistics.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 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)*, pages 1059–1069.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. page 10.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did What: A Large-Scale Person-Centered Cloze Dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, Austin, Texas. Association for Computational Linguistics.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 Task 11: Machine Comprehension Using Commonsense Knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 747–757, New Orleans, Louisiana. Association for Computational Linguistics.

Anselmo Peñas, Christina Unger, and Axel-Cyrille Ngonga Ngomo. 2014. Overview of CLEF Question Answering Track 2014. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 300–306. Springer, Cham.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 12, pages 1532–1543.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765.

Luis Nieto Piña and Richard Johansson. 2015. A Simple and Efficient Method to Generate Word Sense Representations. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 465–472.

Benedicte Pierrejean and Ludovic Tanguy. 2018. Towards Qualitative Word Embeddings Evaluation: Measuring Neighbors Variation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 32–39.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1:8.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. CoQA: A Conversational Question Answering Challenge. *arXiv:1808.07042 [cs]*.

Matthew Richardson, Christopher J C Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA, 18-21 October 2013.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (Too Many) Problems of Analogical Reasoning with Word Vectors. In *Proceedings of the*

6th Joint Conference on Lexical and Computational Semantics (SEM 2017)*, pages 135–148.

Anna Rogers, Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. What's in Your Embedding, And How It Predicts Task Performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2690–2703, Santa Fe, New Mexico, USA, August 20-26, 2018. Association for Computational Linguistics.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

Karl Stratos. 2017. A Sub-Character Architecture for Korean Language Processing. pages 721–726. Association for Computational Linguistics.

Emma Strubell and Andrew McCallum. 2018. Syntax Helps ELMo Understand Semantics: Is Syntax Still Relevant in a Deep Neural Architecture for SRL? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 19–27.

Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2016. Sparse Word Embeddings Using L1 Regularized Online Learning. In *Proceedings of*

the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, pages 2915–2921, New York, New York, USA. AAAI Press.

Yaming Sun, Lei Lin, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. 2014. Radical-Enhanced Chinese Character Embedding. In *Neural Information Processing*, Lecture Notes in Computer Science, pages 279–286. Springer, Cham.

Alon Talmor and Jonathan Berant. 2018. The Web as a Knowledge-Base for Answering Complex Questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. 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)*, pages 4149–4158.

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The importance of being recurrent for modeling hierarchical structure. *arXiv preprint arXiv:1803.03585*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. NewsQA: A Machine Comprehension Dataset. *arXiv:1611.09830 [cs]*.

Yulia Tsvetkov, Manaal Faruqui, and Chris Dyer. 2016. Correlation-based Intrinsic Evaluation of Word Vector Representations. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 111–115.

Peter D. Turney. 2006. Similarity of Semantic Relations. *Comput. Linguist.*, 32(3):379–416.

Jaakko J. Väyrynen and Timo Honkela. 2004. Word category maps based on emergent features created by ICA. *Proceedings of the STeP*, 19:173–185.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *arXiv:1905.00537 [cs]*.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018*

EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2018. Dialogue Natural Language Inference. *arXiv:1811.00671 [cs]*.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors Influencing the Surprising Instability of Word Embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wielfaert, Kris Heylen, Jocelyne Daems, Dirk Speelman, and Dirk Geeraerts. 2014. Towards a Lexicologically Informed Parameter Evaluation of Distributional Modelling in Lexical Semantics.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A Broad-Coverage Challenge Corpus for Sentence Understanding through

Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, 17-21 September 2015. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Mark Yatskar. 2019. A Qualitative Comparison of CoQA, SQuAD 2.0 and QuAC. 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)*, pages 2318–2323.

Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. 2017. Joint Embeddings of Chinese Words, Characters, and Fine-grained Subcharacter Components. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 286–291, Copenhagen, Denmark, September 7–11, 2017. Association for Computational Linguistics.

Zhiguo Yu, Trevor Cohen, Byron Wallace, Elmer Bernstam, and Todd Johnson. 2016. Retrofitting word vectors of mesh terms to improve semantic similarity measures. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 43–51, Austin, Texas, November 5, 2016. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal Common-sense Inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.