Learning Word Representations by Jointly Modeling Syntagmatic and Paradigmatic Relations

Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng

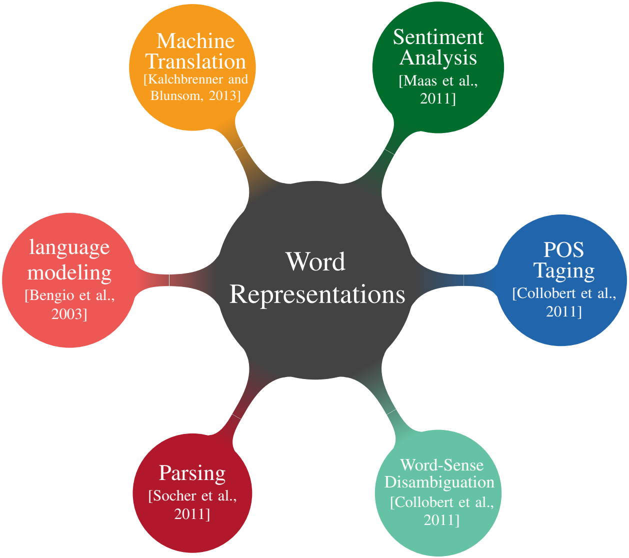
ofey.sunfei@gmail.com

CAS Key Lab of Network Data Science and Technology

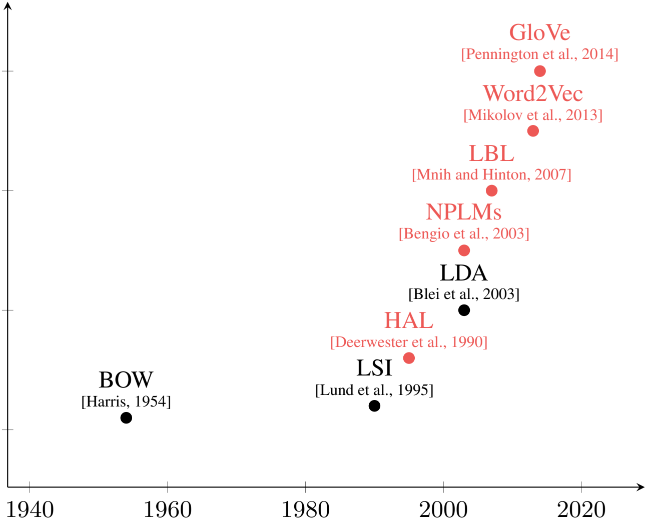
Institute of Computing Technology, Chinese Academy of Sciences

September 25, 2015

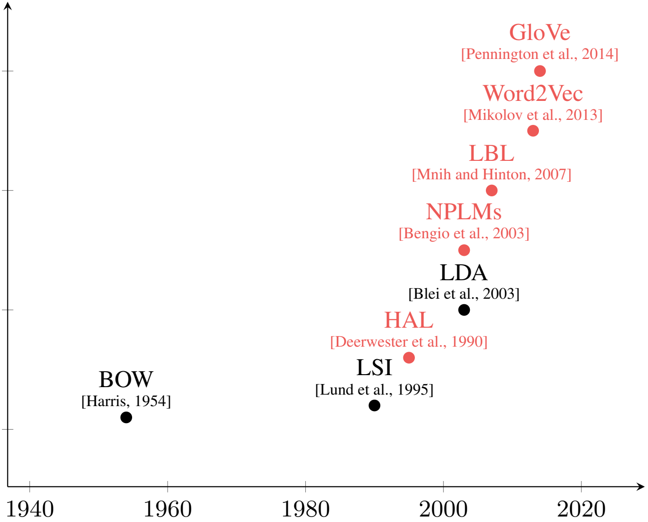
Word Representations



Word Representations Models



Word Representations Models



Relations?

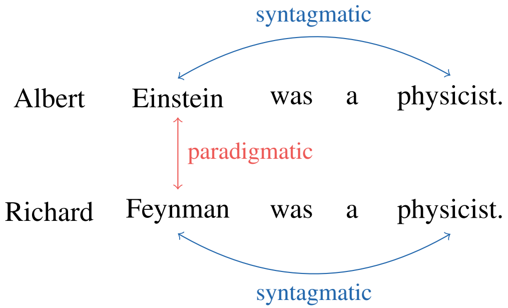
One Hypothesis
Two Interpretation

The Distributional Hypothesis [Harris, 1954, Firth, 1957]

"*You shall know a word by the company it keeps*."

—J.R. Firth

Syntagmatic and Paradigmatic Relations [Gabrilovich and Markovitch, 2007]



- Syntagmatic: words co-occur in the same text region
- Paradigmatic: words occur in the same context, may not at the same time

Syntagmatic

syntagmatic

Albert Einstein was a physicist.

Richard Feynman was a physicist.

syntagmatic

Syntagmatic

syntagmatic

Albert Einstein was a physicist.

Richard Feynman was a physicist.

syntagmatic

	d_1	d_2
Einstein	1	0
Feynman	0	1
physicist	1	1

Syntagmatic

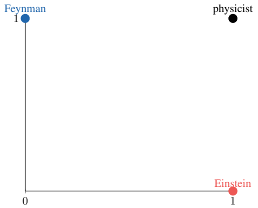
syntagmatic

Albert **Einstein** was a **physicist**.

Richard **Feynman** was a **physicist**.

syntagmatic

	d_1	d_2
Einstein	1	0
Feynman	0	1
physicist	1	1



Syntagmatic

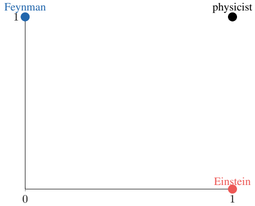
syntagmatic

Albert Einstein was a physicist.

Richard Feynman was a physicist.

syntagmatic

	d_1	d_2
Einstein	1	0
Feynman	0	1
physicist	1	1



LSI, LDA, PV-DBOW $\cdots$

Paradigmatic

Albert	Einstein	was	a	physicist.
	↕ paradigmatic			
Richard	Feynman	was	a	physicist.

Paradigmatic

Albert **Einstein** was a physicist.

$\updownarrow$ paradigmatic

Richard **Feynman** was a physicist.

	Einstein	Feynman	physicist
Einstein	0	0	1
Feynman	0	0	1
physicist	1	1	0

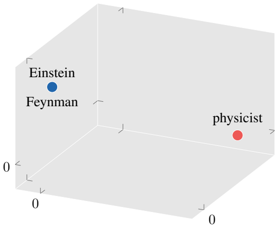
Paradigmatic

Albert **Einstein** was a **physicist**.

$\updownarrow$ paradigmatic

Richard **Feynman** was a **physicist**.

	Einstein	Feynman	physicist
Einstein	0	0	1
Feynman	0	0	1
physicist	1	1	0



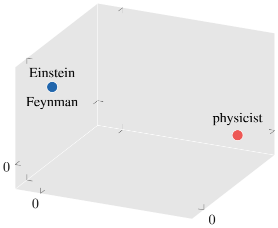
Paradigmatic

Albert Einstein was a physicist.

↕ paradigmatic

Richard Feynman was a physicist.

	Einstein	Feynman	physicist
Einstein	0	0	1
Feynman	0	0	1
physicist	1	1	0



NLMs, Word2Vec, GloVe $\cdots$

Motivation

syntagmatic

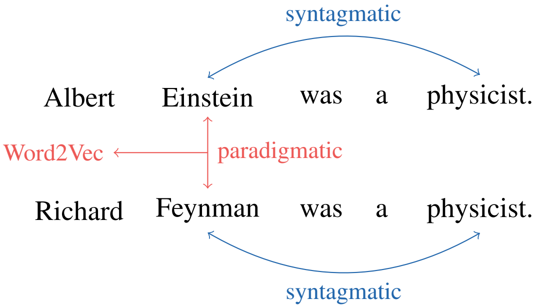
Albert Einstein was a physicist.

paradigmatic

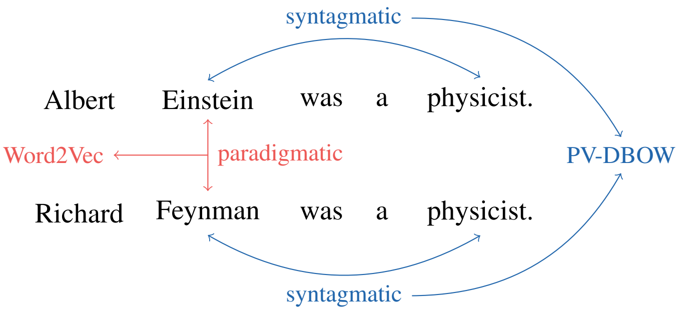
Richard Feynman was a physicist.

syntagmatic

Motivation

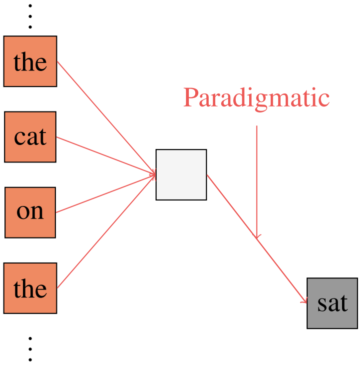

syntagmatic
Albert Einstein was a physicist.
Word2Vec
paradigmatic
Richard Feynman was a physicist.
syntagmatic

Motivation


syntagmatic
Albert Einstein was a physicist.
Word2Vec
paradigmatic
PV-DBOW
Richard Feynman was a physicist.
syntagmatic

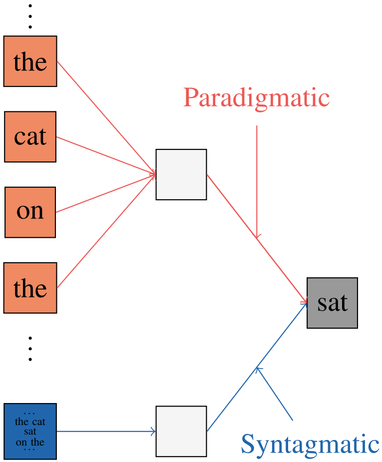
Model

Parallel Document Context Model (PDC)



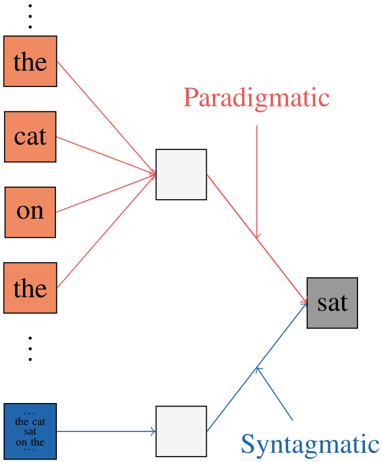
$$\ell = \sum_{n=1}^{N} \sum_{w_i^n \in d_n} \log p(w_i^n | h_i^n)$$

Parallel Document Context Model (PDC)



$$\ell = \sum_{n=1}^{N} \sum_{w_i^n \in d_n} \left(\log p(w_i^n | h_i^n) + \log p(w_i^n | d_n) \right)$$

Parallel Document Context Model (PDC)



$$\ell = \sum_{n=1}^{N} \sum_{w_i^n \in d_n} \left(\log p(w_i^n | h_i^n) + \log p(w_i^n | d_n) \right)$$

Negative Sampling

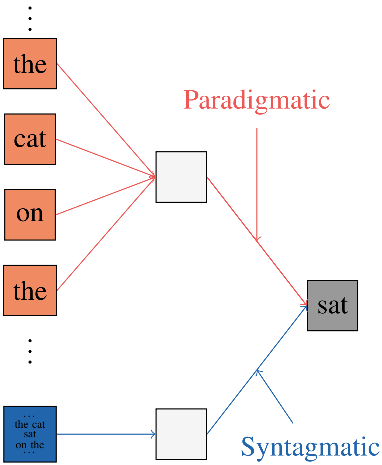
$$\ell = \sum_{n=1}^{N} \sum_{w_i^n \in d_n} \Big(\log \sigma(\vec{w_i^n} \cdot \vec{h_i^n}) + \log \sigma(\vec{w_i^n} \cdot \vec{d_n})$$

$$+ k \cdot \mathbb{E}_{w' \sim P_{\mathrm{nw}}} \log \sigma(-\vec{w'} \cdot \vec{h_i^n})$$

$$+ k \cdot \mathbb{E}_{w' \sim P_{\mathrm{nw}}} \log \sigma(-\vec{w'} \cdot \vec{d_n}) \Big)$$

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

Parallel Document Context Model (PDC)



$$\ell = \sum_{n=1}^{N} \sum_{w_i^n \in d_n} \left(\log p(w_i^n | h_i^n) + \log p(w_i^n | d_n) \right)$$

Negative Sampling

$$\ell = \sum_{n=1}^{N} \sum_{w_i^n \in d_n} \Big(\log \sigma(\vec{w_i^n} \cdot \vec{h_i^n}) + \log \sigma(\vec{w_i^n} \cdot \vec{d_n})$$

$$+ k \cdot \mathbb{E}_{w' \sim P_{\mathrm{nw}}} \log \sigma(-\vec{w'} \cdot \vec{h_i^n})$$

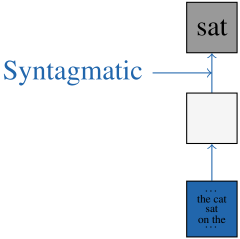
$$+ k \cdot \mathbb{E}_{w' \sim P_{\mathrm{nw}}} \log \sigma(-\vec{w'} \cdot \vec{d_n}) \Big)$$

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

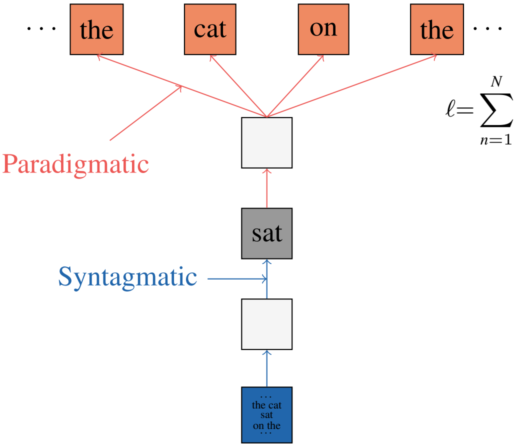
PDC	PV-DM
MF for W-D + W-C [Levy and Goldberg, 2014]	not clear

Hierarchical Document Context Model (HDC)

$$\ell = \sum_{n=1}^{N} \sum_{w_i^n \in d_n} \log p(w_i^n | d_n)$$

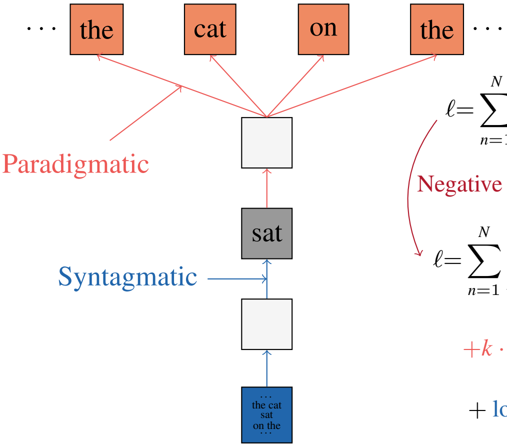


Hierarchical Document Context Model (HDC)



$$\ell = \sum_{n=1}^{N} \sum_{w_i^n \in d_n} \left(\log p(w_i^n | d_n) + \sum_{\substack{j=i-L \\ j \neq i}}^{i+L} \log p(c_j^n | w_i^n) \right)$$

Hierarchical Document Context Model (HDC)



$$\ell = \sum_{n=1}^{N} \sum_{w_i^n \in d_n} \left(\log p(w_i^n | d_n) + \sum_{\substack{j=i-L \\ j \neq i}}^{i+L} \log p(c_j^n | w_i^n) \right)$$

Negative Sampling

$$\ell = \sum_{n=1}^{N} \sum_{w_i^n \in d_n} \Bigg(\sum_{\substack{j=i-L \\ j \neq i}}^{i+L} \Big(\log \sigma(\vec{c_j^n} \cdot \vec{w_i^n}) + k \cdot \mathbb{E}_{c' \sim P_{\text{nc}}} \log \sigma(-\vec{c'} \cdot \vec{w_i^n}) \Big) + \log \sigma(\vec{w_i^n} \cdot \vec{d_n}) + k \cdot \mathbb{E}_{w' \sim P_{\text{nw}}} \log \sigma(-\vec{w'} \cdot \vec{d_n}) \Bigg)$$

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

Relation with Existing Models

- CBOW, SG, PV-DBOW
  - Sub-model
- Global Context-Aware Neural Language Model [Huang et al., 2012]
  - neural language model
  - weighted average of all word vectors

Experiments

Experiments Plan

- Qualitative Evaluations
  - Verify word representations learned by different relations
- Quantitative Evaluations
  - Word Analogy Task
  - Word Similarity Task

Experimental Settings

Corpus:

model	corpus	size
C&W [Collobert et al., 2011]	Wikipedia 2007 + Reuters RCV1	0.85B
HPCA [Lebret and Collobert, 2014]	Wikipedia 2012	1.6B
GloVe	Wikipedia 2014+ Gigaword5	6B
GCANLM, CBOW, SG PV-DBOW, PV-DM, PDC, HDC	Wikipedia 2010	1B

Parameters Setting:

window	negative	iteration	learning rate		noise distribution
10	10	20	0.025[1]	0.05[2]	$\propto \#(w)^{0.75}$

[1] SG, PV-DBOW, HDC [2] CBOW, PV-DM, PDC

Qualitative Evaluations

Top 5 similar words to **Feynman**

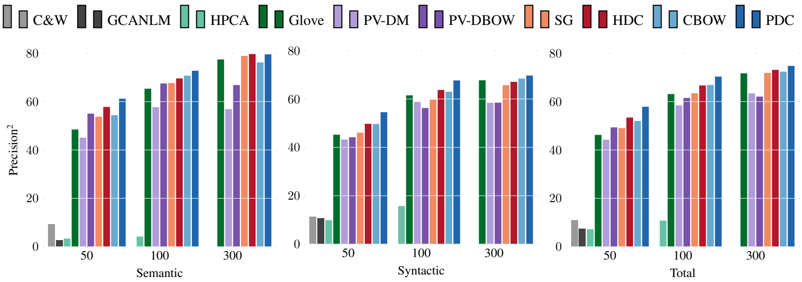
CBOW	SG	PDC	HDC	PV-DBOW
einstein	schwinger	geometrodynamics	schwinger	physicists
schwinger	quantum	bethe	electrodynamics	spacetime
bohm	bethe	semiclassical	bethe	geometrodynamics
bethe	einstein	schwinger	semiclassical	tachyons
relativity	semiclassical	perturbative	quantum	einstein

Paradigmatic Syntagmatic

Word Analogy

- Test Set
  - Google [Mikolov et al., 2013]
    - Semantic: "Beijing is to China as Paris is to _____"
    - Syntactic: "big is to bigger as deep is to _____"
- Solution:
  - $\arg \max_{\substack{x \in W, x \neq a \\ x \neq b,\ x \neq c}} (\vec{b} + \vec{c} - \vec{a}) \cdot \vec{x}$
- Metric:
  - percentage of questions answered correctly

Word Analogy

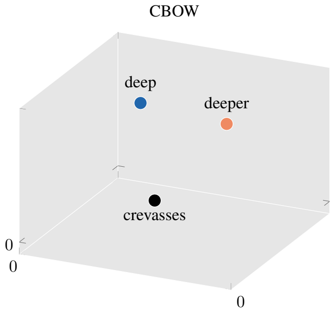


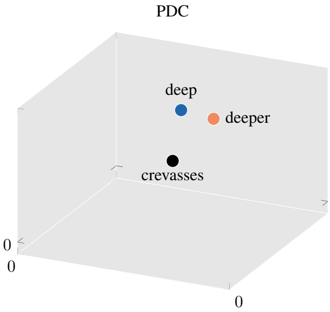
- Word2Vec and GloVe are very strong baselines.
- PDC and HDC outperform CBOW and SG respectively.

[2] percentage of questions answered correctly

Case Study

big: bigger $\sim$ deep: deeper





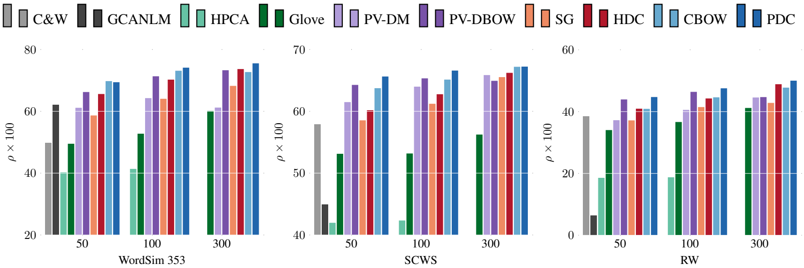
CBOW: shallower ×

PDC: deeper √

Word Similarity

- Test Set
  - WordSim-353 [Finkelstein et al., 2002]
  - Stanford's Contextual Word Similarities (SCWS) [Huang et al., 2012]
  - Rare Word (RW) [Luong et al., 2013]
- Detail:
  - Word pair with similarity score assigned by human
  - (tiger cat 7.35)
- Evaluation Metric:
  - spearman rank correlation

Word Similarity



- PV-DBOW does well.
- PDC and HDC outperform CBOW and SG respectively.

Summary

- Revisit word representation models through syntagmatic and paradigmatic relations.
- Two novel models modeling syntagmatic and paradigmatic relations simultaneously.
- State-of-the-art results.

Thanks

Q & A

More Infomration:

http://ofey.me/projects/wordrep

References I

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003).
A neural probabilistic language model.
J. Mach. Learn. Res., 3:1137–1155.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent dirichlet allocation.
J. Mach. Learn. Res., 3:993–1022.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011).
Natural language processing (almost) from scratch.
J. Mach. Learn. Res., 12:2493–2537.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).
Indexing by latent semantic analysis.
Journal of the American Society for Information Science, 41(6):391–407.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., andGadi Wolfman, Z. S., and Ruppin, E. (2002).
Placing search in context: The concept revisited.
ACM Trans. Inf. Syst., 20(1):116–131.

Firth, J. R. (1957).
A synopsis of linguistic theory 1930-55.
Studies in Linguistic Analysis (special volume of the Philological Society), 1952-59:1–32.

References II

Gabrilovich, E. and Markovitch, S. (2007).
Computing semantic relatedness using wikipedia-based explicit semantic analysis.
In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012).
Improving word representations via global context and multiple word prototypes.
In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kalchbrenner, N. and Blunsom, P. (2013).
Recurrent continuous translation models.
In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Lebret, R. and Collobert, R. (2014).
Word embeddings through hellinger pca.
In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–490. Association for Computational Linguistics.

References III

Lund, K., Burgess, C., and Atchley, R. A. (1995).
Semantic and associative priming in a high-dimensional semantic space.
In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665.

Luong, M.-T., Socher, R., and Manning, C. D. (2013).
Better word representations with recursive neural networks for morphology.
In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113. Association for Computational Linguistics.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011).
Learning word vectors for sentiment analysis.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 142–150, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013).
Efficient estimation of word representations in vector space.
In *Proceedings of Workshop of ICLR*.

Mnih, A. and Hinton, G. (2007).
Three new graphical models for statistical language modelling.
In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 641–648, New York, NY, USA. ACM.

Pennington, J., Socher, R., and Manning, C. D. (2014).
Glove: Global vectors for word representation.
In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

References IV

Socher, R., Lin, C. C., Manning, C., and Ng, A. Y. (2011).
Parsing natural scenes and natural language with recursive neural networks.
In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 129–136, New York, NY, USA. ACM.

Levy, O. and Goldberg Y. (2014).
Neural Word Embedding as Implicit Matrix Factorization.
In Getoor, L. and Scheffer, T., editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185, Montreal, Quebec, Canada. Curran Associates, Inc.