Neural Text Embeddings for Information Retrieval

(WSDM 2017 Tutorial)

Bhaskar Mitra
Microsoft
Cambridge, UK
bmitra@microsoft.com

Nick Craswell
Microsoft
Bellevue, WA, USA
nickcr@microsoft.com

1. MOTIVATIONS

The traditional approach in information retrieval (IR) for estimating relevance of a document to a query is to count the repetitions of the query terms in the contents of the document. The content typically includes the document's body text and the title, but may also include anchor texts linking to the document, and search queries for which the document was previously viewed. In contrast to term based IR, latent semantic models [3, 5, 15] proposed to learn dense vector representations of words and match the document to the query in low-dimensional embedding space. This makes sense, because as noted by Robertson [31], every term in the document contain some information about the document's relevance, irrespective of whether or not the term itself appears in the query.

Over last few years, neural representation learning for text has demonstrated significant improvements in many natural language processing tasks like language modelling [18] and machine translation [2]. While successes from traditional representation learning approaches in IR have been somewhat limited [1], recently there has been a renewed interest in the applications of neural embedding models for retrieval tasks. Much of the recent breakthroughs focus on applications of word embeddings [6, 9, 11, 28], and learning vector representations for short text similarity [17, 35]. Other works have explored using vector representations for document ranking [27], assisting users to formulate rare queries [26], and contextual entity search [10]. Recent tutorials [23] and workshops [4] have explored some aspects of deep learning in the context of IR tasks. This tutorial will focus on the fundamentals of neural representation learning for IR and try to reconcile these new neural embedding models with early representation learning approaches in IR.

2. TOPICS

A tentative schedule and outline of topics is presented below:

09.00–10.30 Early morning session

- Brief history of embeddings in IR: Salton's classic vector space model [34]; latent semantic models [3, 5, 15]; limitations of embeddings models in IR [1, 40]
- Neural word embeddings: Word2vec [24]; explicit vector space representations [22]; GloVe [30]; density-based representations [38]
- Understanding relationships in embedding spaces: different tasks, different embeddings [14, 29, 37]
- Word embeddings for IR: query-document matching [9, 28]; query re-writing [6, 11, 32]; term re-weighting [42]

10:30–11:00 Coffee break

11:00–12:00 Late morning session

- Short and long text embeddings: semantic hashing [33]; Paragraph2vec [21]; DSSM [17, 36]; short text similarity [16, 19, 20, 35]; document ranking [27]
- Other applications in retrieval: query auto-completion [26]; session modelling [25]; user modelling [7]; cross-lingual retrieval [12, 39]; multimedia embeddings [8, 13]
- Tools: CNTK demo [41]

12:00–14:00 Lunch

All materials (including slides presented) will be made available online (http://bit.ly/NeuIRTutorial-WSDM2017) for download. No additional resources or equipments will be necessary.

WSDM 2017 February 06-10, 2017, Cambridge, United Kingdom

ACM ISBN 978-1-4503-4675-7/17/02.

DOI: http://dx.doi.org/10.1145/3018661.3022755

References

[1] A. Atreya and C. Elkan. Latent semantic indexing (lsi) fails for trec collections. *ACM SIGKDD Explorations Newsletter*, 12(2):5–10, 2011.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3: 993–1022, 2003.

[4] N. Craswell, W. B. Croft, J. Guo, B. Mitra, and M. de Rijke. Report on the sigir 2016 workshop on neural information retrieval (neu-ir). 2016.

[5] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[6] F. Diaz, B. Mitra, and N. Craswell. Query expansion with locally-trained word embeddings. In *Proc. ACL*, 2016.

[7] A. M. Elkahky, Y. Song, and X. He. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proc. WWW*, pages 278–288, 2015.

[8] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. *arXiv preprint arXiv:1411.4952*, 2014.

[9] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones. Word embedding based generalized language model for information retrieval. In *Proc. SIGIR*, pages 795–798. ACM, 2015.

[10] J. Gao, P. Pantel, M. Gamon, X. He, L. Deng, and Y. Shen. Modeling interestingness with deep neural networks. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2014.

[11] M. Grbovic, N. Djuric, V. Radosavljevic, and N. Bhamidipati. Search retargeting using directed query embeddings. In *Proc. WWW*, pages 37–38. International World Wide Web Conferences Steering Committee, 2015.

[12] P. Gupta, K. Bali, R. E. Banchs, M. Choudhury, and P. Rosso. Query expansion for mixed-script information retrieval. In *Proc. SIGIR*, pages 677–686. ACM, 2014.

[13] X. He, R. Srivastava, J. Gao, and L. Deng. Joint learning of distributed representations for images and texts. *arXiv preprint arXiv:1504.03083*, 2015.

[14] F. Hill, K. Cho, S. Jean, C. Devin, and Y. Bengio. Not all neural embeddings are born equal. *arXiv preprint arXiv:1410.0718*, 2014.

[15] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. SIGIR*, pages 50–57. ACM, 1999.

[16] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In *Proc. NIPS*, pages 2042–2050, 2014.

[17] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proc. CIKM*, pages 2333–2338. ACM, 2013.

[18] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

[19] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[20] T. Kenter and M. de Rijke. Short text similarity with word embeddings. In *Proc. CIKM*, volume 15, page 115.

[21] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

[22] O. Levy, Y. Goldberg, and I. Ramat-Gan. Linguistic regularities in sparse and explicit word representations. *CoNLL-2014*, page 171, 2014.

[23] H. Li and Z. Lu. Deep learning for information retrieval.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[25] B. Mitra. Exploring session context using distributed representations of queries and reformulations. In *Proc. SIGIR*, pages 3–12. ACM, 2015.

[26] B. Mitra and N. Craswell. Query auto-completion for rare prefixes. In *Proc. CIKM*. ACM, To appear, 2015.

[27] B. Mitra, F. Diaz, and N. Craswell. Learning to match using local and distributed representations of text for web search. *arXiv preprint arXiv:1610.08136*, 2016.

[28] B. Mitra, E. Nalisnick, N. Craswell, and R. Caruana. A dual embedding space model for document ranking. *arXiv preprint arXiv:1602.01137*, 2016.

[29] E. Nalisnick, B. Mitra, N. Craswell, and R. Caruana. Improving document ranking with dual word embeddings. In *Proc. WWW*, 2016.

[30] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. *Proc. EMNLP*, 12: 1532–1543, 2014.

[31] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60 (5):503–520, 2004.

[32] D. Roy, D. Paul, M. Mitra, and U. Garain. Using word embeddings for automatic query expansion. *arXiv preprint arXiv:1606.07608*, 2016.

[33] R. Salakhutdinov and G. Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7): 969–978, 2009.

[34] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11): 613–620, 1975.

[35] A. Severyn and A. Moschitti. Learning to rank short text pairs with convolutional deep neural networks. In *Proc. SIGIR*, pages 373–382. ACM, 2015.

[36] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil. Learning semantic representations using convolutional neural networks for web search. In *Proc. WWW*, pages 373–374, 2014.

[37] F. Sun, J. Guo, Y. Lan, J. Xu, and X. Cheng. Learning word representations by jointly modeling syntagmatic and paradigmatic relations. In *Proc. ACL*, 2015.

[38] L. Vilnis and A. McCallum. Word representations via gaussian embedding. *arXiv preprint arXiv:1412.6623*, 2014.

[39] I. Vulić and M.-F. Moens. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proc. SIGIR*, pages 363–372. ACM, 2015.

[40] X. Yan, J. Guo, S. Liu, X. Cheng, and Y. Wang. Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In *Proceedings of the SIAM International Conference on Data Mining*, 2013.

[41] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang, et al. An introduction to computational networks and the computational network toolkit. Technical report, Tech. Rep. MSR, Microsoft Research, 2014, http://codebox/cntk, 2014.

[42] G. Zheng and J. Callan. Learning to reweight terms with distributed representations. In *Proc. SIGIR*, pages 575–584. ACM, 2015.