Deep Learning for NLP (without Magic)
References

Richard Socher,* Yoshua Bengio,† and Christopher Manning*
*Department of Computer Science, Stanford University
†Department of computer science and operations research, U. Montréal

July 8, 2012
ACL 2012 Tutorial

References

Ando, Rie Kubota and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Machine Learning Research* 6:1817–1853.

Bannard, C. and C. Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL*.

Bengio, Y. 2009. Learning deep architectures for AI. *Foundations & Trends in Mach. Learn.* 2(1):1–127.

Bengio, Yoshua. 2012. Practical recommendations for gradient-based training of deep architectures. Tech. rep., arXiv:1206.5533.

Bengio, Yoshua, Réjean Ducharme, and Pascal Vincent. 2001. A neural probabilistic language model. In T. K. Leen, T. G. Dietterich, and V. Tresp, eds., *Advances in NIPS 13*, pages 932–938. MIT Press.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *J. Machine Learning Research* 3:1137–1155.

Bengio, Y., P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Tr. Neural Networks* 5(2):157–166.

Blitzer, John, Kilian Weinberger, Lawrence Saul, and Fernando Pereira. 2005. Hierarchical distributed representations for statistical language modeling. In *NIPS'2004*. Cambridge, MA: MIT Press.

Bordes, Antoire, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2012. Joint learning of words and meaning representations for open-text semantic parsing. *AISTATS'2012* .

Bordes, Antoire, Jason Weston, Ronan Collobert, and Yoshua Bengio. 2011. Learning structured embeddings of knowledge bases. In *AAAI 2011*.

Bottou, L. 2011. From machine learning to machine reasoning. *CoRR* abs/1102.1808.

Brown, Peter F., Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics* 18(4):467–479.

Clark, Alexander. 2003. Combining distributional and morphological information for part of speech induction˙ In *EACL 2003*, pages 59–66.

Collobert, R. and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML'2008*.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12:2493–2537.

Costa, F., P. Frasconi, V. Lombardo, and G. Soda. 2003. Towards incremental parsing of natural language using recursive neural networks. *Applied Intelligence* 19.

Dahl, George E., Dong Yu, Li Deng, and Alex Acero. 2012. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 20(1):33–42.

Dauphin, Y., X. Glorot, and Y. Bengio. 2011. Large-scale learning of embeddings with reconstruction sampling. In *Proceedings of the 28th International Conference on Machine learning*, ICML '11.

Erhan, Dumitru, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why does unsupervised pre-training help deep learning? *J. Machine Learning Res.* 11:625–660.

Firth, J. R. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pages 1–32. Oxford: Philological Society. Reprinted in F. R. Palmer (ed), *Selected Papers of J. R. Firth 1952–1959*, London: Longman, 1968.

Glorot, Xavier and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS'2010*, pages 249–256.

Glorot, Xavier, Antoire Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *AISTATS'2011*.

Goller, C. and A. Küchler. 1996. Learning task-dependent distributed representations by backpropagation through structure. In *Proceedings of the International Conference on Neural Networks (ICNN-96)*.

Goodfellow, Ian, Quoc Le, Andrew Saxe, and Andrew Ng. 2009. Measuring invariances in deep networks. In *NIPS 22*, pages 646–654.

Gould, S., R. Fulton, and D. Koller. 2009. Decomposing a Scene into Geometric and Semantically Consistent Regions. In *ICCV*.

Gunawardana, A., M. Mahajan, A. Acero, and J. Platt. 2005. Conditional random fields for phone classification. In *Interspeech*, pages 1117–1120. MIT Press.

Hendrickx, I., S.N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz. 2010. Semeval-2010 task 8: Multiway classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*.

Hinton, G. E. 1990. Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence* 46(1-2).

Hinton, Geoffrey. E. 2010. A practical guide to training restricted Boltzmann machines. Tech. Rep. UTML TR 2010-003, Department of Computer Science, University of Toronto.

Huang, Eric H., Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *ACL 2012*.

Koo, Terry, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL*, pages 595–603.

Le, Quoc, Jiquan Ngiam, Adam Coates, Abhik Lahiri, Bobby Prochnow, and Andrew Ng. 2011. On optimization methods for deep learning. In *Proc. ICML'2011*. ACM.

Le, Quoc, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Greg Corrado, Kai Chen, Jeff Dean, and Andrew Ng. 2012. Building high-level features using large scale unsupervised learning. In *ICML'2012*.

Lee, Honglak, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. 2009a. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML'2009*.

Lee, Honglak, Peter Pham, Yan Largman, and Andrew Ng. 2009b. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *NIPS'2009*.

Martin, Sven, Jörg Liermann, and Hermann Ney. 1998. Algorithms for bigram and trigram word clustering. *Speech Communication* 24:19–37.

Menchetti, S., F. Costa, P. Frasconi, and M. Pontil. 2005. Wide coverage natural language processing using kernel methods and neural networks for structured data. *Pattern Recognition Letters* 26(12).

Mikolov, Tomas, Anoop Deoras, Stefan Kombrink, Lukas Burget, and Jan Cernocky. 2011. Empirical evaluation and combination of advanced language modeling techniques. In *Proc. 12th annual conference of the international speech communication association (INTERSPEECH 2011)*.

Mnih, Andriy and Geoffrey E. Hinton. 2007. Three new graphical models for statistical language modelling. In *ICML'2007*, pages 641–648.

Mnih, Andriy and Geoffrey E. Hinton. 2009. A scalable hierarchical distributed language model. In *NIPS 21*, pages 1081–1088.

Morin, Frédéric and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In *AISTATS'2005*, pages 246–252.

Pollack, J. B. 1990. Recursive distributed representations. *Artificial Intelligence* 46.

Quattoni, Ariadna, Michael Collins, and Trevor Darrell. 2005. Conditional random fields for object recognition. In *NIPS'2004*, pages 1097–1104. MIT Press.

rahman Mohamed, Abdel, George Dahl, and Geoffrey Hinton. 2012. Acoustic modeling using deep belief networks. *IEEE Trans. on Audio, Speech and Language Processing* 20(1):14–22.

Ratliff, N., J. A. Bagnell, and M. Zinkevich. 2007. (Online) subgradient methods for structured prediction. In *Eleventh International Conference on Artificial Intelligence and Statistics (AIStats)*.

Reisinger, Joseph and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *HLT-NAACL 2010*, pages 109–117.

Rifai, Salah, Yann Dauphin, Pascal Vincent, and Yoshua Bengio. 2012. A generative process for contractive auto-encoders. In *ICML'2012*.

Rifai, Salah, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. 2011. Contracting auto-encoders: Explicit invariance during feature extraction. In *ICML'2011*.

Rink, B. and S. Harabagiu. 2010. UTD: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*.

Schwenk, Holger. 2007. Continuous space language models. *Computer speech and language* 21:492–518.

Schwenk, H. and J-L. Gauvain. 2002. Connectionist language modeling for large vocabulary continuous speech recognition. In *ICASSP*, pages 765–768. Orlando, Florida.

Schwenk, Holger, Anthony Rousseau, and Mohammed Attik. 2012. Large, pruned or continuous space language models on a gpu for statistical machine translation. In *Workshop on the future of language modeling for HLT*.

Seide, Frank, Gang Li, and Dong Yu. 2011. Conversational speech transcription using context-dependent deep neural networks. In *Interspeech 2011*, pages 437–440.

Sha, Fei and Fernando C. N. Pereira. 2003. Shallow parsing with conditional random fields. In *HLT-NAACL*.

Smith, Noah A. and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 354–362. Ann Arbor, Michigan: Association for Computational Linguistics.

Socher, Richard, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011a. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems 24*.

Socher, Richard, Cliff C. Lin, Andrew Y. Ng, and Christopher D. Manning. 2011b. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*.

Socher, R., C. D. Manning, and A. Y. Ng. 2010. Learning continuous phrase representations and syntactic parsing with recursive neural networks. In *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*.

Socher, Richard, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011c. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, pages 252–259.

Turian, Joseph, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proc. ACL'2010*, pages 384–394. Association for Computational Linguistics.

Vincent, Pascal, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *ICML 2008*, pages 1096–1103.