

## University of Groningen

### One Model to Rule them All

Bjerva, Johannes

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*

Publisher's PDF, also known as Version of record

*Publication date:*  
2017

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

Bjerva, J. (2017). *One Model to Rule them All: multitask and Multilingual Modelling for Lexical Analysis*. [Thesis fully internal (DIV), University of Groningen]. Rijksuniversiteit Groningen.

**Copyright**

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

**Take-down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

*Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.*

# Bibliography

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I. J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D. G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P. A., Vanhoucke, V., Vasudevan, V., Viégas, F. B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.
- Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., van Noord, R., Ludmann, P., Nguyen, D.-D., and Bos, J. (2017). The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations. In *EACL*, pages 242–247.
- Agić, Ž., Johannsen, A., Plank, B., Alonso, H. M., Schluter, N., and Søgaard, A. (2016). Multilingual projection for parsing truly low-resource languages. *TACL*, 4:301–312.
- Agić, Ž., Plank, B., and Søgaard, A. (2017). Cross-lingual tagger evaluation without test data. In *EACL*, pages 248–253.
- Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., and Wiebe, J. (2015). Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *SemEval 2015*, pages 252–263.

- Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. (2014). Semeval-2014 task 10: Multilingual semantic textual similarity. In *SemEval*, pages 81–91.
- Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval*, pages 497–511.
- Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *SemEval*, pages 385–393.
- Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. (2013). \*SEM 2013 shared task: Semantic textual similarity. In *\*SEM-SemEval*, pages 32–43.
- Agirre, E., Cer, D., Diab, M., Lopez-Gazpio, I., and Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *SemEval*.
- Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual nlp. *CoNLL-2013*.
- Aldarmaki, H. and Diab, M. (2016). GWU NLP at SemEval-2016 Shared Task 1: Matrix factorization for crosslingual STS. In *SemEval*, pages 663–667.
- Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. (2016). Many languages, one parser. *TACL*, 4:431–444.
- Ando, R. K. and Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853.
- Argyriou, A., Evgeniou, T., and Pontil, M. (2007). Multi-task feature learning. In *Advances in neural information processing systems*, pages 41–48.
- Arora, S., Bhaskara, A., Ge, R., and Ma, T. (2014). Provable bounds for learning some deep representations. In *ICML*, pages 584–592.
- Arroyo-Fernández, I. and Ruiz, I. V. M. (2017). Lipn-iimas at semeval-2017 task 1: Subword embeddings, attention recurrent neural networks and cross word alignment for semantic textual similarity. In *SemEval*.

- Ataman, D., de Souza, J. G. C., Turchi, M., and Negri, M. (2016). Fbk hlt-mt at semeval-2016 task 1: Cross-lingual semantic similarity measurement using quality estimation features and compositional bilingual word embeddings. In *SemEval*, pages 570–576.
- Augenstein, I. and Søgaard, A. (2017). Multi-task learning of keyphrase boundary classification. In *ACL*, pages 341–346.
- Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate.
- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *ACL-COLING*, pages 86–90, Université de Montréal, Montreal, Quebec, Canada.
- Ballesteros, M., Dyer, C., and Smith, N. A. (2015). Improved transition-based parsing by modeling characters instead of words with lstms. In *EMNLP*, pages 349–359.
- Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, volume 1.
- Barzdins, G. and Gosko, D. (2016). Riga: Impact of smatch extensions and character-level neural translation on amr parsing accuracy. In *SemEval*.
- Baxter, J. (1997). A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39.
- Baxter, J. et al. (2000). A model of inductive bias learning. *J. Artif. Intell. Res. (JAIR)*, 12(149-198):3.
- Beltagy, I., Roller, S., Cheng, P., Erk, K., and Mooney, R. J. (2016). Representing meaning with a combination of logical and distributional models. *Computational Linguistics*.
- Bengio, Y., Boulanger-Lewandowski, N., and Pascanu, R. (2013). Advances in optimizing recurrent networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8624–8628. IEEE.

- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.
- Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., and Lin, Z. (2015). Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*.
- Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.
- Benton, A., Mitchell, M., and Hovy, D. (2017). Multitask learning for mental health conditions with limited social media data. In *EACL*, volume 1, pages 152–162.
- Berant, J. and Liang, P. (2014). Semantic parsing via paraphrasing. In *ACL*, pages 1415–1425.
- Bingel, J. and Søggaard, A. (2017). Identifying beneficial task relations for multi-task learning in deep neural networks. In *EACL*, pages 164–169.
- Bjerva, J. (2014). Multi-class animacy classification with semantic features. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–75.
- Bjerva, J. (2016). Byte-based language identification with deep convolutional networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 119–125.
- Bjerva, J. (2017a). Quantifying the Effects of Multilinguality in NLP Sequence Prediction Tasks. *Under review*.
- Bjerva, J. (2017b). Will my auxiliary tagging task help? Estimating Auxiliary Tasks Effectivity in Multi-Task Learning. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden*, number 131, pages 216–220. Linköping University Electronic Press, Linköpings universitet. Best short paper.
- Bjerva, J. and Börstell, C. (2016). Morphological Complexity Influences Verb-Object Order in Swedish Sign Language. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 137–141.

- Bjerva, J., Bos, J., and Haagsma, H. (2016a). The Meaning Factory at SemEval-2016 Task 8: Producing AMRs with Boxer. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1179–1184.
- Bjerva, J., Bos, J., Van der Goot, R., and Nissim, M. (2014). The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 642–646, Dublin, Ireland.
- Bjerva, J., Grigonytė, G., Östling, R., and Plank, B. (2017). Neural networks and spelling features for native language identification. In *Proceedings of Shared Task on NLI at BEA17*.
- Bjerva, J. and Östling, R. (2017a). Cross-lingual Learning of Semantic Textual Similarity with Multilingual Word Representations. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden*, number 131, pages 211–215. Linköping University Electronic Press, Linköpings universitet.
- Bjerva, J. and Östling, R. (2017b). Multilingual word representations for semantic textual similarity. In *Proceedings of SemEval 2017: International Workshop on Semantic Evaluation*.
- Bjerva, J., Plank, B., and Bos, J. (2016b). Semantic tagging with deep residual networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3531–3541.
- Bjerva, J. and Praet, R. (2015). Word embeddings pointing the way for late antiquity. In *9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities (LaTeCH 2015)*, pages 53–57.
- Bjerva, J. and Praet, R. (2016). Rethinking intertextuality through a word-space and social network approach – the case of Cassiodorus. *Journal of Data Mining and Digital Humanities*, *accepted, in revision*.
- Bos, J. (2008). Wide-Coverage Semantic Analysis with Boxer. In Bos, J. and Delmonte, R., editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 277–286. College Publications.

- Bos, J. (2014). Semantic annotation issues in parallel meaning banking. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 17.
- Bos, J., Basile, V., Evang, K., Venhuizen, N. J., and Bjerva, J. (2017). *The Groningen Meaning Bank*, pages 463–496. Springer Netherlands, Dordrecht.
- Bos, J. and Nissim, M. (2015). Uncovering noun-noun compound relations by gamification. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NoDaLiDa 2015, May 11-13, 2015, Vilnius, Lithuania*, number 109, pages 251–255. Linköping University Electronic Press.
- Bottou, L. (1998). Online algorithms and stochastic approximations. In Saad, D., editor, *Online Learning*. Cambridge University Press, Cambridge, UK.
- Brady, M. L., Raghavan, R., and Slawny, J. (1989). Back propagation fails to separate where perceptrons succeed. *IEEE Transactions on Circuits and Systems*, 36(5):665–674.
- Brants, T. (2000). Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231.
- Busger op Vollenbroek, M., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., and Nissim, M. (2016). Gron-UP: Groningen user profiling. In *Proceedings of CLEF 2016*.
- Butler, A. (2010). *The Semantics of Grammatical Dependencies*, volume 23. Emerald Group Publishing Limited.
- Caruana, R. (1997). Multitask learning. *Machine Learning*, 28 (1):41–75.
- Caruana, R. A. (1993). Multitask connectionist learning. In *In Proceedings of the 1993 Connectionist Models Summer School*. Citeseer.
- Chamberlain, J. (2014). Groupsourcing: Distributed problem solving using social networks. In *Second AAAI Conference on Human Computation and Crowdsourcing*.
- Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750.

- Cheng, H., Fang, H., and Ostendorf, M. (2015). Open-domain name error detection using a multi-task rnn. In *EMNLP*.
- Cho, K. (2015). Natural language understanding with distributed representation. *arXiv preprint arXiv:1511.07916*.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.
- Chollet, F. (2015). Keras. <https://github.com/fchollet/keras>.
- Chomsky, N. (2005). Three factors in language design. *Linguistic inquiry*, 36(1):1–22.
- Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2015). The loss surfaces of multilayer networks. In *AISTATS*.
- Chrupała, G. (2013). Text segmentation with character-level text embeddings. In *Workshop on Deep Learning for Audio, Speech and Language Processing, ICML*.
- Chung, J., Cho, K., and Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. *ACL*.
- Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
- Ciaramita, M. and Altun, Y. (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *EMNLP*, pages 594–602.
- Collins, J., Sohl-Dickstein, J., and Sussillo, D. (2017). Capacity and trainability in recurrent neural networks.
- Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.



- Conneau, A., Schwenk, H., Barrault, L., and Lecun, Y. (2016). Very deep convolutional networks for natural language processing. *arXiv preprint arXiv:1606.01781*.
- Copestake, A., Flickinger, D., Sag, I., and Pollard, C. (2005). Minimal recursion semantics: An introduction. *Journal of Research on Language and Computation*, 3(2-3):281–332.
- Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Kübler, S., Yarowsky, D., Eisner, J., and Hulden, M. (2017). The CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *CoNLL-SIGMORPHON*.
- Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., and Hulden, M. (2016). The sigmorphon 2016 shared task—morphological reinflection. In *ACL 2016*, page 10.
- Coulmance, J., Marty, J.-M., Wenzek, G., and Benhalloum, A. (2015). Trans-gram, fast cross-lingual word-embeddings. In *EMNLP*, pages 1109–1113, Lisbon, Portugal.
- Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.
- Curran, J., Clark, S., and Bos, J. (2007). Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *ACL*, pages 33–36.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314.
- Das, D., Chen, D., Martins, A. F., Schneider, N., and Smith, N. A. (2014). Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.
- Daumé III, H. (2009). Bayesian multitask learning with latent hierarchies. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 135–142. AUAI Press.
- Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941.

- Deng, L. and Wiebe, J. (2015). Mpqa 3.0: An entity/event-level sentiment corpus. In *HLT-NAACL*, pages 1323–1328.
- dos Santos, C. N. and Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. In *ICML*, pages 1818–1826.
- Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.
- Duong, L., Cohn, T., Bird, S., and Cook, P. (2015). Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *ACL*, pages 845–850.
- Dyer, C., Ballesteros, M., Ling, W., Matthews, A., and Smith, N. A. (2015). Transition-based dependency parsing with stack long short-term memory. In *ACL-IJCNLP*, pages 334–343, Beijing, China.
- Dyer, C., Kuncoro, A., Ballesteros, M., and Smith, N. A. (2016). Recurrent neural network grammars. In *NAACL*, pages 199–209.
- Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.
- Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.
- Evang, K. (2016). *Cross-lingual Semantic Parsing with Categorical Grammars*. PhD thesis, University of Groningen.
- Evang, K., Basile, V., Chrupała, G., and Bos, J. (2013). Elephant: Sequence labeling for word and sentence segmentation. In *EMNLP*, pages 1422–1426.
- Evang, K. and Bos, J. (2016). Cross-lingual learning of an open-domain semantic parser. In *COLING*.
- Evgeniou, T. and Pontil, M. (2004). Regularized multi-task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM.
- Fang, M. and Cohn, T. (2017). Model transfer for tagging low-resource languages using a bilingual dictionary. In *ACL*, pages 587–593.

- Faruqui, Y. T. S. S. M. and Lample, G. (2016). Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *NAACL-HLT*, pages 1357–1366.
- Firth, J. R. (1957). A synopsis of linguistic theory. pages 1930–1955. 1952–1959:1–32.
- Gal, Y. and Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027.
- Gallistel, C. (2016). The neurobiological bases for the computational theory of mind. *Perspectives on the Work of Jerry Fodor*. L. Gleitman and R. G. d. Almeida, Editors. New York, Oxford University Press.
- Gallistel, C. R. and King, A. P. (2011). *Memory and the computational brain: Why cognitive science will transform neuroscience*, volume 6. John Wiley & Sons.
- Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The paraphrase database. In *NAACL-HLT*, pages 758–764.
- Georgi, R., Xia, F., and Lewis, W. (2010). Comparing language similarity across genetic and typologically-based groupings. In *COLING*, pages 385–393.
- Gillick, D., Brunk, C., Vinyals, O., and Subramanya, A. (2015). Multilingual language processing from bytes. In *NAACL-HLT*.
- Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256.
- Goldberg, Y. (2015). A primer on neural network models for natural language processing. *arXiv preprint arXiv:1510.00726*.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

- Goodfellow, I. J., Vinyals, O., and Saxe, A. M. (2015). Qualitatively characterizing neural network optimization problems. In *International Conference on Learning Representations*.
- Gori, M. and Tesi, A. (1992). On the problem of local minima in backpropagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(1):76–86.
- Gouws, S. and Søgaard, A. (2015). Simple task-specific bilingual word embeddings. In *HLT-NAACL*, pages 1386–1390.
- Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610.
- Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2016). Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*.
- Guo, J., Che, W., Wang, H., and Liu, T. (2014). Learning sense-specific word embeddings by exploiting bilingual resources. In *COLING*, pages 497–507.
- Guo, J., Che, W., Yarowsky, D., Wang, H., and Liu, T. (2016). A representation learning framework for multi-source transfer parsing. In *Proc. of AAAI*.
- Haagsma, H. and Bjerva, J. (2016). Detecting novel metaphor using selectional preference information. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 10–17.
- Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., and Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951.
- Hahnloser, R. H. and Seung, H. S. (2001). Permitted and forbidden sets in symmetric threshold-linear networks. In *Advances in Neural Information Processing Systems*, pages 217–223.
- Harris, Z. (1954). Distributional structure. *Word*, 10:146—162.
- Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579.

- He, K., Zhang, X., Ren, S., and Sun, J. (2015a). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015b). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027*.
- Hebb, D. O. (1949). *The organization of behavior: A neuropsychological approach*. John Wiley & Sons.
- Hermann, K. M., Das, D., Weston, J., and Ganchev, K. (2014). Semantic frame identification with distributed word representations. In *ACL*, pages 1448–1458.
- Hill, F., Cho, K., and Korhonen, A. (2016). Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*.
- Hinton, G. E. (2012). Lecture 6.5 - rmsprop, coursera: Neural networks for machine learning. Technical report.
- Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.
- Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591.
- Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology (London)*, 160(1):106–154.

- Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology (London)*, 195(1):215–243.
- Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(03):311–325.
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.
- Iyyer, M., Manjunatha, V., Boyd-Graber, J. L., and Daumé III, H. (2015). Deep unordered composition rivals syntactic methods for text classification. In *ACL*, pages 1681–1691.
- Jirenghed, D.-A., Rasmussen, A., Johansson, F., and Hesslow, G. (2017). Learned response sequences in cerebellar purkinje cells. *Proceedings of the National Academy of Sciences*, 114(23):6127–6132.
- Johansson, F., Jirenghed, D.-A., Rasmussen, A., Zucca, R., and Hesslow, G. (2014). Memory trace and timing mechanism localized to cerebellar Purkinje cells. *Proceedings of the National Academy of Sciences*, 111(41):14930–14934.
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2016). Google’s multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.
- Jordan, M. I. (1997). Serial order: A parallel distributed processing approach. *Advances in psychology*, 121:471–495.
- Jozefowicz, R., Zaremba, W., and Sutskever, I. (2015). An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2342–2350.
- Jurgens, D. and Navigli, R. (2014). It’s all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *TACL*, 2:449–464.
- Kanerva, P., Kristoferson, J., and Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the Cognitive Science Society*, volume 1.

- Kang, Z., Grauman, K., and Sha, F. (2011). Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 521–528.
- Kann, K. and Schütze, H. (2016). Med: The lmu system for the sigmorphon 2016 shared task on morphological inflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 62–70, Berlin, Germany.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.
- Klementiev, A., Titov, I., and Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In *COLING*, pages 1459–1474, Mumbai, India.
- Klerke, S., Goldberg, Y., and Søgaard, A. (2016). Improving sentence compression by learning to predict gaze. *arXiv preprint arXiv:1604.03357*.
- Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit.*, Phuket, Thailand.
- Konstas, I., Iyer, S., Yatskar, M., Choi, Y., and Zettlemoyer, L. (2017). Neural amr: Sequence-to-sequence models for parsing and generation. *arXiv preprint arXiv:1704.08381*.
- Korchagina, N. (2017). Normalizing medieval german texts: from rules to deep learning. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, number 133, pages 12–17. Linköping University Electronic Press, Linköpings universitet.
- Krogh, A. and Hertz, J. A. (1992). A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957.

- Kulmizev, A., Blankers, B., Bjerva, J., Nissim, M., van Noord, G., Plank, B., and Wieling, M. (2017). The power of character n-grams in native language identification. In *Proceedings of Shared Task on NLI at BEA17*.
- Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.
- Lample, G. and Chaplot, D. S. (2017). Playing fps games with deep reinforcement learning. In *AAAI*, pages 2140–2146.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998a). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R. (1998b). Efficient backprop. In *Neural Networks: Tricks of the Trade*, pages 9–50. Springer.
- LeCun, Y. et al. (1989). Generalization and network design strategies. *Connectionism in perspective*, pages 143–155.
- Levy, O. and Goldberg, Y. (2014a). Dependency-based word embeddings. In *ACL*, pages 302–308.
- Levy, O. and Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.
- Lo, C.-k., Goutte, C., and Simard, M. (2016). Cnrc at semeval-2016 task 1: Experiments in crosslingual semantic textual similarity. *SemEval*, pages 668–673.
- Loeff, N. and Farhadi, A. (2008). Scene discovery by matrix factorization. In *European Conference on Computer Vision*, pages 451–464. Springer.
- Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2015). Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.
- Malaviya, C., Neubig, G., and Littell, P. (2017). Learning language representations for typology prediction. *arXiv preprint arXiv:1707.09569*.



- Manning, C. D. (2015). Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.
- Marcus, M., Santorini, B., and Marcinkiewicz, M. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. (2014). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *SemEval*, pages 1–8.
- Martínez Alonso, H. and Plank, B. (2017). When is multitask learning effective? semantic sequence prediction under varying data conditions. In *EACL*, pages 44–53.
- Maurer, A. (2006). Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7(Jan):117–139.
- McDonald, R., Petrov, S., and Hall, K. (2011a). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of EMNLP*, pages 62–72.
- McDonald, R., Petrov, S., and Hall, K. (2011b). Multi-source transfer of delexicalized dependency parsers. In *EMNLP*, pages 62–72.
- Melis, G., Dyer, C., and Blunsom, P. (2017). On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.
- Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *NAACL-HLT*, pages 746–751.
- Miller, G. A., Leacock, C., Teng, R., and Bunker, R. T. (1993). A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308.

- Minsky, M. and Papert, S. (1988). *Perceptrons: an introduction to computational geometry (expanded edition)*. MIT Press, Cambridge, Ma.
- Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L., and Jin, Z. (2016). How transferable are neural networks in nlp applications? *arXiv preprint arXiv:1603.06111*.
- Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814.
- Nivre, J., Agić, Ž., Ahrenberg, L., Aranzabe, M. J., Asahara, M., Atutxa, A., Ballesteros, M., Bauer, J., Bengoetxea, K., Bhat, R. A., Bick, E., Bosco, C., Bouma, G., Bowman, S., Candito, M., Cebiroğlu Eryiğit, G., Celano, G. G. A., Chalub, F., Choi, J., Çöltekin, Ç., Connor, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eli, M., Erjavec, T., Farkas, R., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Ginter, F., Goenaga, I., Gojenola, K., Gökürmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Habash, N., Hajič, J., Hà My, L., Haug, D., Hladká, B., Hohle, P., Ion, R., Irimia, E., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kanayama, H., Kanerva, J., Kotsyba, N., Krek, S., Laippala, V., Lê Hong, P., Lenci, A., Ljubešić, N., Lyashevskaya, O., Lynn, T., Makazhanov, A., Manning, C., Mărănduc, C., Mareček, D., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Missilä, A., Mititelu, V., Miyao, Y., Montemagni, S., More, A., Mori, S., Moskalevskiy, B., Muischnek, K., Mustafina, N., Müürisepp, K., Nguyen Thi, L., Nguyen Thi Minh, H., Nikolaev, V., Nurmi, H., Ojala, S., Osenova, P., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Plank, B., Popel, M., Pretkalniņa, L., Prokopidis, P., Puolakainen, T., Pyysalo, S., Rademaker, A., Ramasamy, L., Real, L., Rituma, L., Rosa, R., Saleh, S., Sanguinetti, M., Saulīte, B., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shakurova, L., Shen, M., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Tanaka, T., Tsarfaty, R., Tyers, F., Uematsu, S., Uria, L., van Noord, G., Varga, V., Vincze, V., Washington, J. N., Žabokrtský, Z., Zeldes, A., Zeman, D., and Zhu, H. (2017). Universal dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

- Nivre, J., Agić, Ž., Aranzabe, M. J., Asahara, M., Atutxa, A., Ballesteros, M., Bauer, J., Bengoetxea, K., Bhat, R. A., Bosco, C., Bowman, S., Celano, G. G. A., Connor, M., de Marneffe, M.-C., Diaz de Ilarraza, A., Dobrovoljc, K., Dozat, T., Erjavec, T., Farkas, R., Foster, J., Galbraith, D., Ginter, F., Goenaga, I., Gojenola, K., Goldberg, Y., Gonzales, B., Guillaume, B., Hajič, J., Haug, D., Ion, R., Irimia, E., Johannsen, A., Kanayama, H., Kanerva, J., Krek, S., Laippala, V., Lenci, A., Ljubešić, N., Lynn, T., Manning, C., Mărănduc, C., Mareček, D., Martínez Alonso, H., Mašek, J., Matsumoto, Y., McDonald, R., Missilä, A., Mititelu, V., Miyao, Y., Montemagni, S., Mori, S., Nurmi, H., Osenova, P., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Petrov, S., Piitulainen, J., Plank, B., Popel, M., Prokopidis, P., Pyysalo, S., Ramasamy, L., Rosa, R., Saleh, S., Schuster, S., Seeker, W., Seraji, M., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Simov, K., Smith, A., Štěpánek, J., Suhr, A., Szántó, Z., Tanaka, T., Tsarfaty, R., Uematsu, S., Uria, L., Varga, V., Vincze, V., Žabokrtský, Z., Zeman, D., and Zhu, H. (2015). Universal dependencies 1.2. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.
- Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016a). Universal dependencies v1: A multilingual treebank collection. In *LREC*.
- Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016b). Universal dependencies v1: A multilingual treebank collection. In *LREC*, pages 1659–1666.
- Obozinski, G., Taskar, B., and Jordan, M. I. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252.
- Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.
- Östling, R. (2014). Bayesian word alignment for massively parallel texts. In *EACL*, pages 123–127.
- Östling, R. (2015). *Bayesian models for multilingual word alignment*. PhD thesis, Department of Linguistics, Stockholm University.

- Östling, R. (2016). Morphological reinflection with convolutional neural networks. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.
- Östling, R. and Bjerva, J. (2017). SU-RUG at the CoNLL-SIGMOR-PHON 2017 shared task: Morphological Inflection with Attentional sequence-to-sequence models. In *Proceedings of the 2017 Meeting of SIGMORPHON*, Vancouver, Canada. Association for Computational Linguistics.
- Östling, R. and Tiedemann, J. (2016). Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*, 106(1):125–146.
- Östling, R. and Tiedemann, J. (2017). Continuous multilinguality with language vectors. In *EACL*, pages 644–649.
- Ouchi, H., Duh, K., and Matsumoto, Y. (2014). Improving dependency parsers with supertags. In *EACL*, pages 154–158.
- Ouchi, H., Duh, K., and Matsumoto, Y. (2016). Transition-based dependency parsing exploiting supertags. In *IEEE/ACM Transactions on Audio, Speech and Language Processing*, volume 24.
- Palatucci, M., Pomerleau, D., Hinton, G. E., and Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, pages 1410–1418.
- Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *ICML (3)*, 28:1310–1318.
- Plank, B. (2016). Keystroke dynamics as signal for shallow syntactic parsing. In *COLING*, pages 609–619.
- Plank, B., Søgaard, A., and Goldberg, Y. (2016). Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *ACL*.
- Pradhan, S. S., Ward, W., Hacioglu, K., Martin, J. H., and Jurafsky, D. (2004). Shallow semantic parsing using support vector machines. In *HLT-NAACL*, pages 233–240.

- Quattoni, A., Collins, M., and Darrell, T. (2008). Transfer learning for image classification with sparse prototype representations. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.
- Rama, T. and Borin, L. (2015). Comparative evaluation of string similarity measures for automatic language classification. In *Quantitative Linguistics*, volume 69. De Gruyter Mouton.
- Rasooli, M. and Collins, M. (2015). Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of EMNLP*.
- Rosa, R., Zeman, D., Mareček, D., and Žabokrtský, Z. (2017). Slavic forest, norwegian wood. pages 210–219.
- Rosenblatt, F. (1957). The perceptron: a perceiving and recognizing automaton. Technical report, Cornell Aeronautical Laboratory, inc.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638.
- Ruder, S., Bingel, J., Augenstein, I., and Søgaard, A. (2017). Sluice networks: Learning what to share between loosely related tasks. *arXiv preprint arXiv:1705.08142*.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.
- Sahlgren, M. (2005). An introduction to random indexing. In *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE*, volume 5.
- Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.
- Saxe, A. M., McClelland, J. L., and Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *ICLR*.
- Schaul, T., Antonoglou, I., and Silver, D. (2014). Unit tests for stochastic optimization. In *International Conference on Learning Representations*.

- Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.
- Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S., and van Gerven, M. (2017). Cnn-based encoding and decoding of visual object recognition in space and time. *bioRxiv*, page 118091.
- Semeniuta, S., Severyn, A., and Barth, E. (2016). Recurrent dropout without memory loss. *arXiv preprint arXiv:1603.05118*.
- Shannon, C. E. and Weaver, W. (1949). *The mathematical theory of communication*. Urbana, Ill.: University of Illinois press, 1964.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.
- Sjons, J., Hörberg, T., Östling, R., and Bjerva, J. (2017). Articulation rate in swedish child-directed speech increases as a function of the age of the child even when surprisal is controlled for. In *Proceedings of Interspeech 2017*, Stockholm, Sweden.
- Socher, R., Chen, D., Manning, C. D., and Ng, A. (2013a). Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pages 926–934.
- Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. (2013b). Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943.
- Søgaard, A. and Goldberg, Y. (2016). Deep multi-task learning with low level tasks supervised at lower layers. In *ACL*, volume 2, pages 231–235.
- Søgaard, A., Johannsen, A., Plank, B., Hovy, D., and Martinez, H. (2014). What’s in a p-value in nlp? In *Proceedings of the eighteenth conference on computational natural language learning (CONLL’14)*, pages 1–10.
- Sontag, E. D. and Sussmann, H. J. (1989). Backpropagation can give rise to spurious local minima even for networks without hidden layers. *Complex Systems*, 3(1):91–106.

- Specia, L., Shah, K., De Souza, J. G., and Cohn, T. (2013). Quest-a translation quality estimation framework. In *ACL (Conference System Demonstrations)*, pages 79–84.
- Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.
- Straka, M., Hajic, J., and Straková, J. (2016). Ud-pipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *LREC*.
- Sundermeyer, M., Schlüter, R., and Ney, H. (2012). Lstm neural networks for language modeling. In *Interspeech*, pages 194–197.
- Šuster, S. (2016). *Empirical studies on word representations*. PhD thesis, University of Groningen.
- Sutskever, I., Martens, J., and Hinton, G. E. (2011). Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.
- Täckström, O. (2013). *Predicting Linguistic Structure with Incomplete and Cross-Lingual Supervision*. PhD thesis, PhD Thesis - Uppsala University.
- Täckström, O., Das, D., Petrov, S., McDonald, R., and Nivre, J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. *TACL*, 1:1–12.
- Täckström, O., McDonald, R., and Uszkoreit, J. (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 477–487.

- Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *ACL-IJCNLP*, pages 1556–1566.
- Thibodeau, P. H. and Boroditsky, L. (2013). Natural language metaphors covertly influence reasoning. *PLoS one*, 8(1):e52961.
- Thrun, S. and O’Sullivan, J. (1995). Clustering learning tasks and the selective cross-task transfer of knowledge. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.
- Thrun, S. and Pratt, L. (1998). *Learning to Learn*. Kluwer Academic Publishers, Norwell, MA, USA.
- Tian, J., Zhou, Z., Lan, M., and Wu, Y. (2017). ECNU at SemEval-2017 Task 1: Leverage kernelbased traditional nlp features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In *SemEval*.
- Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *LREC*. European Language Resources Association (ELRA).
- Tiedemann, J. (2014). Rediscovering annotation projection for cross-lingual parser induction. In *COLING*, pages 1854–1864.
- Tiedemann, J. (2015). Cross-lingual dependency parsing with universal dependencies and predicted pos labels. *Depling 2015*, page 340.
- Tiedemann, J., Agić, Ž., and Nivre, J. (2014). Treebank translation for cross-lingual parser induction. In *Proceedings of CoNLL*.
- Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147.
- Tokui, S., Oono, K., Hido, S., and Clayton, J. (2015). Chainer: a next-generation open source framework for deep learning. In *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*.



- Torrallba, A., Murphy, K. P., and Freeman, W. T. (2007). Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5).
- Toutanova, K., Haghighi, A., and Manning, C. D. (2005). Joint learning improves semantic role labeling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 589–596.
- Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *ACL*, pages 384–394.
- van Noord, R. and Bos, J. (2017a). Neural semantic parsing by character-based translation: Experiments with abstract meaning representations. *arXiv preprint arXiv:1705.09980*.
- van Noord, R. and Bos, J. (2017b). The Meaning Factory at SemEval-2017 Task 9: Producing AMRs with Neural Semantic Parsing. *arXiv preprint arXiv:1704.02156*.
- Vaswani, A., Zhao, Y., Fossum, V., and Chiang, D. (2013). Decoding with large-scale neural language models improves translation. In *EMNLP*, pages 1387–1392.
- Veit, A., Wilber, M. J., and Belongie, S. (2016). Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*, pages 550–558.
- Venhuizen, N., Basile, V., Evang, K., and Bos, J. (2013). Gamification for word sense labeling. In *Proc. 10th International Conference on Computational Semantics (IWCS-2013)*, pages 397–403.
- Vilares, D., Gómez-Rodríguez, C., Alonso, M. A., and LyS, G. (2016). One model, two languages: training bilingual parsers with harmonized treebanks. In *ACL*, page 425.
- Vinyals, O., Kaiser, Ł., Koo, T., Petrov, S., Sutskever, I., and Hinton, G. (2015). Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, pages 2773–2781.
- Vossen, P., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., Roventini, A., Bertagna, F., Alonge, A., and Peters, W. (1998). The eurowordnet base concepts and top ontology. *Deliverable D017 D*, 34:D036.

- Šuster, S., Titov, I., and van Noord, G. (2016). Bilingual learning of multi-sense embeddings with discrete autoencoders. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1346–1356, San Diego, California.
- Waibel, A., Hanazawa, T., Hinton, G. E., Shikano, K., and Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339.
- Wang, H., Raj, B., and Xing, E. P. (2017). On the origin of deep learning. *arXiv preprint arXiv:1702.07800*.
- Wang, P., Qian, Y., Soong, F. K., He, L., and Zhao, H. (2015). A unified tagging solution: Bidirectional lstm recurrent neural network with word embedding. *arXiv preprint arXiv:1511.00215*.
- Wichmann, S., Holman, E. W., and Brown (eds.), C. H. (2016). The ASJP Database (version 17).
- Wiener, N. (1948). Cybernetics. *Scientific American*, 179(5):14–19.
- Wolf, L., Hanani, Y., Bar, K., and Dershowitz, N. (2014). Joint word2vec networks for bilingual semantic representations. In *Proceedings of CICLING 2014*.
- Wu, H. (2009). Global stability analysis of a general class of discontinuous neural networks with linear growth activation functions. *Information Sciences*, 179(19):3432–3441.
- Wu, H., Huang, H., Jian, P., Guo, Y., and Su, C. (2017). BIT at SemEval-2017 Task 1: Using semantic information space to evaluate semantic textual similarity. In *SemEval*.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Yang, Z., Salakhutdinov, R., and Cohen, W. (2016). Multi-Task Cross-Lingual Sequence Tagging from Scratch. *arXiv preprint arXiv:1603.06270*.

- Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8.
- Yi, X., Li, R., and Sun, M. (2016). Generating chinese classical poems with rnn encoder-decoder. *arXiv preprint arXiv:1604.01537*.
- Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. In *Proceedings of the 27th British Machine Vision Conference (BMVC)*, *arXiv preprint arXiv:1605.07146*.
- Zeman, D. and Resnik, P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP Workshop on NLP for Less Privileged Languages*.
- Zhang, D., Yuan, B., Wang, D., and Liu, R. (2015a). Joint semantic relevance learning with text data and graph knowledge. In *ACL-IJCNLP 2015*, page 32.
- Zhang, X., Zhao, J., and LeCun, Y. (2015b). Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657.
- Zhou, Y. and Chellappa, R. (1988). Computation of optical flow using a neural network. In *IEEE International Conference on Neural Networks*, volume 1998, pages 71–78.
- Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *LREC*, Paris, France. European Language Resources Association (ELRA).