References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019. doi: 10.18653/v1/N19-1423 pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[3] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith, "Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping," 2020.

[4] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019. doi: 10.18653/v1/P19-1355 pp. 3645–3650. [Online]. Available: https://www.aclweb.org/anthology/P19-1355

[5] F. Olsson, "Bootstrapping named entity annotation by means of active machine learning: A method for creating corpora," Ph.D. dissertation, , SICS, 2008. [Online]. Available: http://spraakdata.gu.se/publikationer/datalinguistica/DL21.pdf

[6] A. Kirsch, J. van Amersfoort, and Y. Gal, "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and

R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/95323660ed2124450caaac2c46b5ed90-Paper.pdf

[7] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 1183–1192. [Online]. Available: http://proceedings.mlr.press/v70/gal17a.html

[8] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," in *Eighth International Conference on Learning Representations (ICLR)*, April 2020. [Online]. Available: https://www.microsoft.com/en-us/research/publication/deep-batch-active-learning-by-diverse-uncertain-gradient-lower-bounds/

[10] M. Yuan, H.-T. Lin, and J. Boyd-Graber, "Cold-start active learning through self-supervised language modeling," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020. doi: 10.18653/v1/2020.emnlp-main.637 pp. 7935–7948. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.637

[11] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, pp. 107–116, 04 1998. doi: 10.1142/S0218488598000094

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735

[13] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997. doi: 10.1109/78.650093

[14] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13. Red Hook, NY, USA: Curran Associates Inc., 2013, p. 3111–3119.

[15] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient estimation of word representations in vector space," 01 2013, pp. 1–12.

[16] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014. doi: 10.3115/v1/D14-1162 pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162

[17] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. of NAACL*, 2018.

[18] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018. doi: 10.18653/v1/P18-1031 pp. 328–339. [Online]. Available: https://www.aclweb.org/anthology/P18-1031

[19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. doi: 10.1109/CVPR.2015.7298935 pp. 3156–3164.

[20] J. Ba, J. Kiros, and G. Hinton, "Layer normalization," 07 2016.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html

[22] J. Alammar, 2018. [Online]. Available: https://jalammar.github.io/illustrated-transformer/

[23] A. Radford and K. Narasimhan, “Improving language understanding by generative pre-training,” 2018.

[24] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, “Supervised learning of universal sentence representations from natural language inference data,” 09 2017. doi: 10.18653/v1/D17-1070 pp. 670–680.

[25] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” 01 2019. doi: 10.18653/v1/D19-1410 pp. 3973–3983.

[26] F. Carlsson, A. C. Gyllensten, E. Gogoulou, E. Y. Hellqvist, and M. Sahlgren, “Semantic re-tuning with contrastive tension,” in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=Ov_sMNau-PF

[27] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, “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*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018. doi: 10.18653/v1/W18-5446 pp. 353–355. [Online]. Available: https://www.aclweb.org/anthology/W18-5446

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019.

[29] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

[30] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “Electra: Pre-training text encoders as discriminators rather than generators,” *ArXiv*, vol. abs/2003.10555, 2020.

[31] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS

[32] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

[33] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," 2020.

[34] M. Malmsten, L. Börjeson, and C. Haffenden, "Playing with words at the national library of sweden – making a swedish bert," 2020.

[35] L. Ramshaw and M. Marcus, "Text chunking using transformation-based learning," in *Third Workshop on Very Large Corpora*, 1995. [Online]. Available: https://www.aclweb.org/anthology/W95-0107

[36] E. F. Tjong Kim Sang and F. De Meulder, "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*, 2003, pp. 142–147. [Online]. Available: https://www.aclweb.org/anthology/W03-0419

[37] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[38] C. A. Thompson, M. E. Califf, and R. J. Mooney, "Active learning for natural language parsing and information extraction," in *Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)*, Bled, Slovenia, June 1999, pp. 406–414. [Online]. Available: http://www.cs.utexas.edu/users/ai-lab?thompson:ml99

[39] G. Tur, R. Schapire, and D. Hakkani-Tur, "Active learning for spoken language understanding," vol. 1, 05 2003. doi: 10.1109/ICASSP.2003.1198771. ISBN 0-7803-7663-3 pp. I–276.

[40] G. Tur, D. Hakkani-Tür, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, no. 2, pp. 171–186, 2005. doi: https://doi.org/10.1016/j.specom.2004.08.002.

[Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639304000962

[41] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., vol. 20. Curran Associates, Inc., 2008. [Online]. Available: https://proceedings.neurips.cc/paper/2007/file/a1519de5b5d44b31a01de013b9b51a80-Paper.pdf

[42] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," ser. EMNLP '08. USA: Association for Computational Linguistics, 2008, p. 1070–1079.

[43] S. Peshterliev, J. Kearney, A. Jagannatha, I. Kiss, and S. Matsoukas, "Active learning for new domains in natural language understanding," 01 2019. doi: 10.18653/v1/N19-2012 pp. 90–96.

[44] D. Shen, J. Zhang, J. Su, G. Zhou, and C.-L. Tan, "Multi-criteria-based active learning for named entity recognition," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, Barcelona, Spain, Jul. 2004. doi: 10.3115/1218955.1219030 pp. 589–596. [Online]. Available: https://www.aclweb.org/anthology/P04-1075

[45] A. Siddhant and Z. Lipton, "Deep bayesian active learning for natural language processing: Results of a large-scale empirical study," 01 2018. doi: 10.18653/v1/D18-1318 pp. 2904–2909.

[46] Y. Shen, H. Yun, Z. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017. doi: 10.18653/v1/W17-2630 pp. 252–256. [Online]. Available: https://www.aclweb.org/anthology/W17-2630

[47] Y. Chen, T. A. Lasko, Q. Mei, J. C. Denny, and H. Xu, "A study of active learning methods for named entity recognition in clinical text," *Journal of Biomedical Informatics*, vol. 58, pp. 11–18, 2015. doi: https://doi.org/10.1016/j.jbi.2015.09.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046415002038

[48] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. ISBN 1558607781 p. 282–289.

[49] Q. Wei, Y. Chen, M. Salimi, J. Denny, Q. Mei, T. Lasko, Q. Chen, S. Wu, A. Franklin, T. Cohen, and H. Xu, "Cost-aware active learning for named entity recognition in clinical text," *Journal of the American Medical Informatics Association : JAMIA*, 2019.

[50] L. Ein-Dor, A. Halfon, A. Gera, E. Shnarch, L. Dankin, L. Choshen, M. Danilevsky, R. Aharonov, Y. Katz, and N. Slonim, "Active Learning for BERT: An Empirical Study," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020. doi: 10.18653/v1/2020.emnlp-main.638 pp. 7949–7962. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.638

[51] D. Grießhaber, J. Maucher, and N. T. Vu, "Fine-tuning BERT for low-resource natural language understanding via active learning," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020. doi: 10.18653/v1/2020.coling-main.100 pp. 1158–1171. [Online]. Available: https://www.aclweb.org/anthology/2020.coling-main.100

[52] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 1321–1330. [Online]. Available: http://proceedings.mlr.press/v70/guo17a.html

[53] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=H1aIuk-RW

[54] W.-N. Hsu and H.-T. Lin, "Active learning by learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, Feb.

2015. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/9597

[55] L. Ahrenberg, J. Frid, and L.-J. Olsson, "A new resource for swedish named-entity recognition," 2020. [Online]. Available: https://gubox.app.box.com/v/SLTC-2020-paper-17

[56] Arbetsförmedlingen AI-Center. AF-BERT. [Online]. Available: https://github.com/af-ai-center/SweBERT

[57] F. Stollenwerk, N. Fastlund, A. Nyqvist, and J. Öhman, "Annotated job ads using swedish language models and named entity recognition," in preparation.

[58] F. Stollenwerk. nerblackbox: a python package to fine-tune transformer-based language models for named entity recognition. [Online]. Available: https://af-ai-center.github.io/nerblackbox/

[59] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[60] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[61] M. Mosbach, M. Andriushchenko, and D. Klakow, "On the stability of fine-tuning {bert}: Misconceptions, explanations, and strong baselines," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=nzpLWnVAyah

[62] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, and Y. Artzi, "Revisiting few-sample {bert} fine-tuning," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=cO1IH43yUF

[63] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 01 2020. doi: 10.18653/v1/2020.acl-main.747 pp. 8440–8451.

[64] K. Margatina, L. Barrault, and N. Aletras, "Bayesian active learning with pretrained language models," 2021.

Appendix A

Hyperparameter Comparison

A.1 ALPS Inspired vs Early Stopping

While the hyperparameters from ALPS may push the performance slightly for full-sized datasets with learning rate decay, it does not allow model training to convergence due to the low number of epochs. Instead, training until convergence using early stopping and a maximum of 50 epochs allows convergence also for early iterations and small datasets. This could be treated as a trade-off. Nevertheless, stable and fair training throughout all iterations is in this thesis deemed more important than perfect fine-tuning for AL experiments. A comparison of the hyperparameters from ALPS and the hyperparameters used in the experiments is shown in Figure A.1. Indeed, models in early iterations perform poorly with a low number of epochs. Furthermore, the variance with early stopping is reduced.

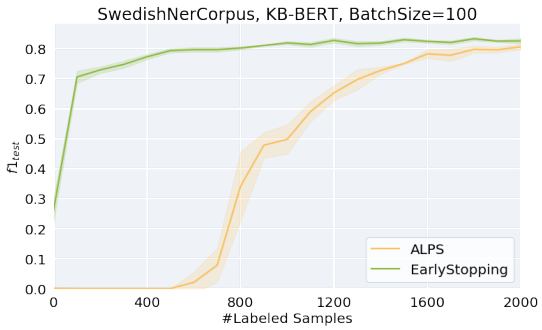

Figure A.1 – Model hyperparameter performance for an increasing number of data labels. The x-axis shows the number of labeled samples beyond the initial seed dataset of 50 samples.