Project Proposal

Thesis Title: Concise and Engaging Title

by

Bas Diender
(2711178)

First Supervisor: Name and Surname
Daily Supervisor: Name and Surname
Second Reader: Name and Surname

Project Proposal

Bas Diender

Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

1 Introduction

Language models like BERT [5] have become a cornerstone technology in NLP research. In short, they are trained for next-word prediction in large datasets of text and, in doing so, pick up on linguistic properties that can be leveraged in a wide variety of downstream language-related tasks. Multilingual language models (MLMs) are trained on texts in many different languages. In the process, they pick up on both linguistic properties that are shared across all languages, as well as language-specific properties shared by only a subset of all languages. As a consequence, an MLM can bootstrap knowledge obtained from one language when dealing with another language in a process called **cross-lingual transfer**. This is particularly advantageous for low-resource languages. Such languages may not have enough data for a language model to adequately pick up on the language's patterns; however, in a multilingual setting, the model can use the patterns it picks up from more well-resourced languages in its training data.

Some linguistic patterns are language-universal, such as the presence of a subject-predicate structure to express who is doing what; whereas others are present only in a subset of the world's languages, such as strictly positioning a subject in front of its predicate. The study of which languages exhibit which patterns is called **linguistic typology**. [4] It is intuitive that an MLM's ability to leverage properties of one language to improve its understanding of another depends on the typological similarity between the two languages. Given the importance of cross-lingual transfer for improving performance on low-resource languages, it is important to subject this intuition to scrutiny. Hence, the goal of this project is to investigate the role of typology in an MLM's adaptability to other languages. In particular, it aims to assess the following hypotheses:

- **H1:** When adapting to a typologically similar language, changes to the model parameters are less extensive than when the model adapts to a more distant language;
- **H2:** When adapting to a typologically similar language, changes to the model parameters are minimal in the layers responsible for handling syntactic features;
- **H3:** The performance of an MLM in a few-shot setting is positively correlated with the typological similarity to the languages it has already adapted to.

The corresponding research questions are as follows:

- **Q1:** How is typological distance reflected in the model parameters? (H1, H2)
- **Q2:** How does typological distance affect cross-lingual transfer, as reflected by downstream performance? (H3)

2 Related Work

The link between typological similarity and successful cross-lingual transfer is well established. Pires et al. [12] found that typological similarity correlates positively with zero-shot learning performance. While they attributed their results to lexical overlap between the source and target languages, Karthikeyan et al. [7] found the same results when the source and target languages had no lexical overlap at all. Instead, they demonstrated structural similarity between two languages to be of greater importance.

Typological relatedness has informed cross-lingual research in low-resource settings, too. Nooralahzadeh et al. [11] use a multi-stage approach to cross-lingual transfer with MAML where between the initial stage of fine-tuning on a large dataset to instill task-specific knowledge and few-shot learning to low-resource languages, they fine-tune on auxiliary languages to optimize the model parameters for quick adaptation. Choenni et al. [1] adapt this method and match auxiliary languages and low-resource languages based on typological similarity to further facilitate cross-lingual transfer.

Tenney et al. [13] set out to quantify where specific types of linguistic information are encoded in BERT, and found each consecutive layer to handle incrementally high-level linguistic information. That is, earlier layers handle concrete features like parts-of-speech tagging, whereas more abstract information like dependency relations and coreference resolution is dealt with in later layers. In the multilingual setting, Choenni & Shutova [2] found that M-BERT's internal representations have a clear typological organization, with the model encoding typological features explicitly and jointly across languages.

This joint encoding could be what underlies the relation between typological similarity and successful cross-lingual transfer. Fine-tuning an MLM on any given language could see it increase the importance of the encodings of the typological features of that language. As a result, after fine-tuning, the model's parameters would be close to what they would have been if the model had been fine-tuned to a typologically similar language.

3 Methodology

The methodology follows Choenni et al. [1] in adapting a three-stage fine-tuning setup outlined in Section 3.2, with the model being fine-tuned for dependency parsing in a variety of languages. The choice for dependency parsing was based on the availability of uniform datasets in a wide variety of languages through Universal Dependencies (UD) [10]. The selection of languages is clarified in Section 3.3, the specifics of the model are discussed in Section 3.1.

3.1 Model

Following Choenni et al., [1] the model used is derived from UDify, [8] which is itself based on M-BERT [5]. UDify incorporates a layer attention mechanism that scales each layer for its importance to the task at hand. That is, during fine-tuning, a weighted sum $\mathbf{e}_j$ is computed for input token j over all layers $i \in [1, ..., 12]$ as follows:

$$\mathbf{e}_j = \sum_i \mathbf{U}_{i,j} \cdot \text{softmax}(\mathbf{w})_i$$

where $\mathbf{U}_{i,j}$ is the output of layer i at token position j and $\mathbf{w}$ is a vector trained alongside the model such that $\mathbf{w}_i$ reflects the importance of layer i. Dependency arcs between tokens are scored using a biaffine attention classifier, [6] after which the optimal parse tree is decoded using the Chu-Liu/Egmonds algorithm. [3]

3.2 Fine-tuning setup

As stated, the fine-tuning setup follows three stages adapted from Choenni et al. [1]

In the first stage, the model M is fine-tuned on a German-language (*deu*) treebank to instill general knowledge of the task at hand. In the second stage, four copies of M are fine-tuned separately on treebanks of auxiliary languages $\ell \in \{nldsweces, hun\}$[1] to obtain four instances of M^ℓ. Finally, each M^ℓ is fine-tuned on each low-resource language $\lambda \in \{gsw, fao, hsb, vep\}$,[2] resulting in 16 instances of $M^{\ell,\lambda}$. If time permits, both the second and third stage are ideally repeated across at least five random seeds.

3.3 Languages

Broadly speaking, language selection is based on their typological similarity and their prevalence in M-BERT's pre-training data as reported by Wu & Dredze. [15] Typological similarity is defined using the syntactic feature-vectors (`syntax_knn`) from the URIEL knowledge base. [9] In particular, each language is represented as a feature-vector, and similarity between languages is taken to be the cosine similarity between their respective vectors.

German was selected as one of the most prominent languages in M-BERT's pre-training data. Moreover, Turc et al. [14] found German (alongside Russian) to be the most effective language for cross-lingual transfer. The auxiliary languages all have a roughly similar amount of data in M-BERT's pre-training data. Their selection is motivated phylogenetically, as they each represent a different level of relatedness to German (Hungarian is unrelated, Czech is in the same family, Swedish is in the same branch, Dutch is in the same subbranch). However, a hierarchical clustering analysis of the `syntax_knn` vectors (Appendix A) found their selection to by typologically valid, too.

4 Experimental Setup

The goal of the project is twofold: Firstly, it aims to show how typology relates to model parameter updates in the fine-tuning stage; secondly, it aims to investigate the effect of typological distance on the success of cross-lingual transfer. Section 4.1 shows how the setup outlined in Section 3.2 serves to achieve these goals. Section 4.2 lists the parameters used, Table B in Appendix B lists the specific datasets used in the experiments.

[1] Dutch, Swedish, Czech, and Hungarian.
[2] Swiss German, Faroese, Upper Sorbian, and Veps.

4.1 Analyses

Q1 In the second stage, four model instances optimized for dependency parsing in German is fine-tuned to do dependency parsing in languages with varying levels of similarity to German. Intuitively, the German-language data in the preceding stage would have prepared it for dependency parsing in the closely related Dutch more than in the unrelated Hungarian. H1 would suggest that the model instance that fine-tunes to Dutch sees less extensive parameter udpates than the one that fine-tunes to Hungarian. In addition, H2 suggests that the updates should primarily be to the earlier layers of the model for Dutch, whereas for Hungarian, they would be across all layers.

The fine-tuning setup includes learned scalar weights $\mathbf{w}_i$ that indicate the importance of each layer i. Following Tenney et al. [13], these can be used to detect where model updates took place for each ℓ. With $\Delta\mathbf{w}^{\ell} = \mathbf{w}^{\ell} - \mathbf{w}^{deu}$ being the shift in layer importance after fine-tuning on ℓ, the expectation is that $\Delta\mathbf{w}_i^{nld}$ is greater for layers i that deal with lexical information (i.e., the earlier ones [13]), but lower for layers i that deal with higher-level typological features, since those are generally shared between Dutch and German. On the other hand, the values of $\Delta\mathbf{w}^{hun}$ are expected to be more uniform, since Hungarian differs significantly from German both lexically and syntactically.

Q2 In the third stage, the model instances for languages ℓ are fine-tuned further to each the low-resource languages λ. Each $M^{\ell,\lambda}$ is then evaluated on the test set for λ, with performance measured as the Labeled Attachment Score (LAS). H3 suggests that greater cosine similarity $s(\ell, \lambda)$ should correspond to higher performance on λ's test set.

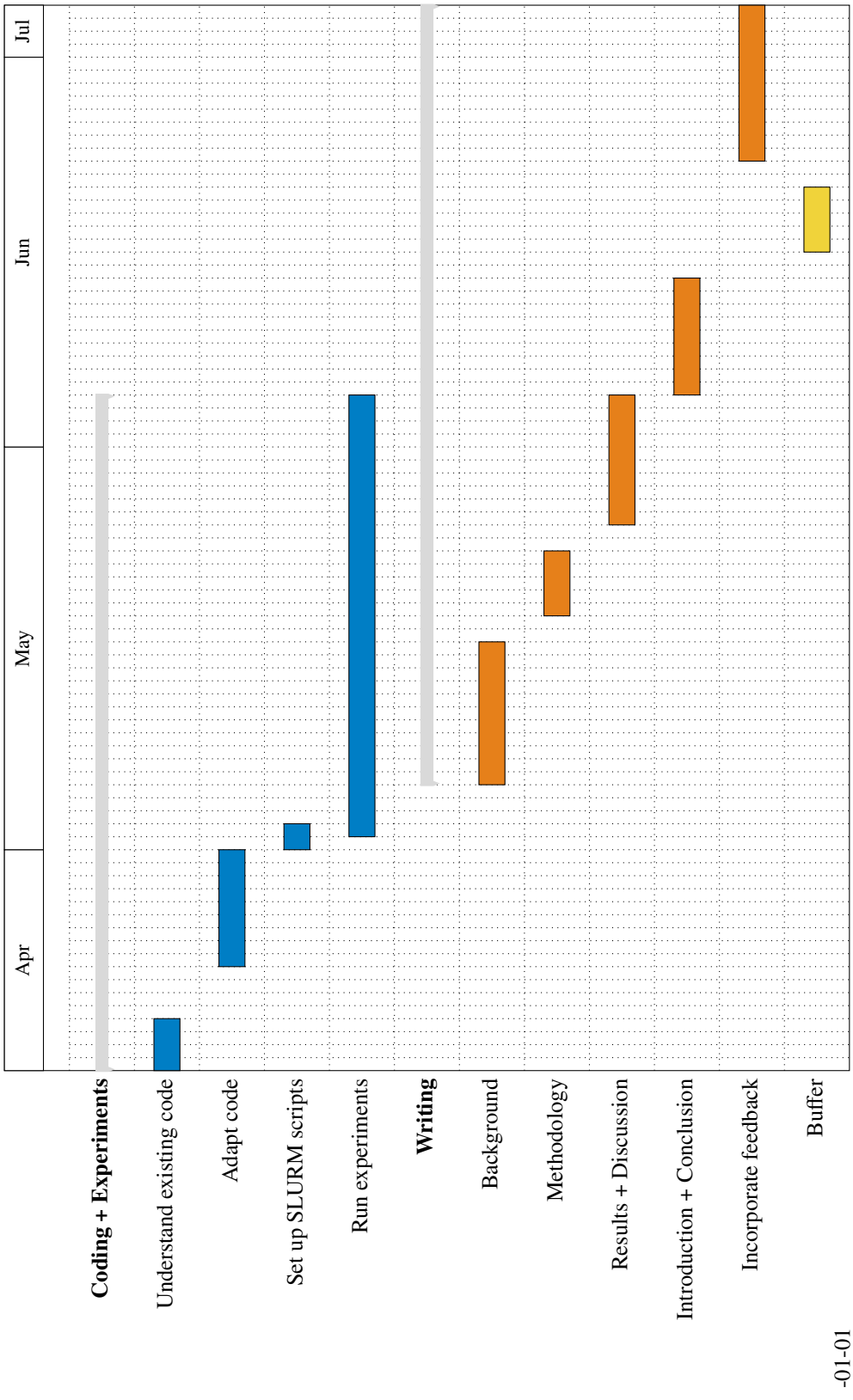
4.2 Parameters

The model is first fine-tuned to Part A[3] of the German training set. Choenni et al. [1] use the English EWT treebank with $\tilde{1}$2.5k samples in this stage and train for 60 epochs. To match the number of examples in this setup, training on the German treebank spans 11 epochs. In the second stage, the model instances are fine-tuned over 1 000 iterations with a batch size of 20 for all languages ℓ. In the third stage, the models are fine-tuned on 20 examples sampled randomly from each treebank of language λ. For the Upper Sorbian treebank, these are sampled from the train set. The other three treebanks do not split their data into separate sets; therefore, the sampled examples are removed from the dataset so as to prevent evaluating on them.

Following Choenni et al. [1], a cosine-based learning rate scheduler with 10 % warm-up and the Adam optimizer are used. For the language model, a learning rate of $1e-04$ is used, and for the classifier, a learning rate of $1e-03$ is used.

[3] The HDT treebank training data is split into various parts. With Part A being larger than the entire English-language treebank used by Choenni et al. [1], the decision was made to use just that part for training.

5 Planning


Apr
May
Jun
Jul
Coding + Experiments
Understand existing code
Adapt code
Set up SLURM scripts
Run experiments
Writing
Background
Methodology
Results + Discussion
Introduction + Conclusion
Incorporate feedback
Buffer
-01-01

References

1. Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing. *Computational Linguistics*, pages 613–641, September 2023.
2. Rochelle Choenni and Ekaterina Shutova. Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology. *Computational Linguistics*, 48(3):635–672, September 2022.
3. Yoeng-Jin Chu. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400, 1965.
4. William Croft. *Typology and Universals*. Cambridge University Press, 2002.
5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
6. Timothy Dozat and Christopher D Manning. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*, 2016.
7. Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual BERT: an empirical study. *CoRR*, abs/1912.07840, 2019.
8. Dan Kondratyuk and Milan Straka. 75 languages, 1 model: Parsing Universal Dependencies universally. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China, November 2019. Association for Computational Linguistics.
9. Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain, April 2017. Association for Computational Linguistics.
10. Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).
11. Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. Zero-shot cross-lingual transfer with meta learning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online, November 2020. Association for Computational Linguistics.
12. Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics.

13. Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics.
14. Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. Revisiting the primacy of english in zero-shot cross-lingual transfer. *arXiv e-prints*, pages arXiv–2106, 2021.
15. Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In Spandana Gella, Johannes Welbl, Marek Rei, Fabio Petroni, Patrick Lewis, Emma Strubell, Minjoon Seo, and Hannaneh Hajishirzi, editors, *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online, July 2020. Association for Computational Linguistics.

A Appendix

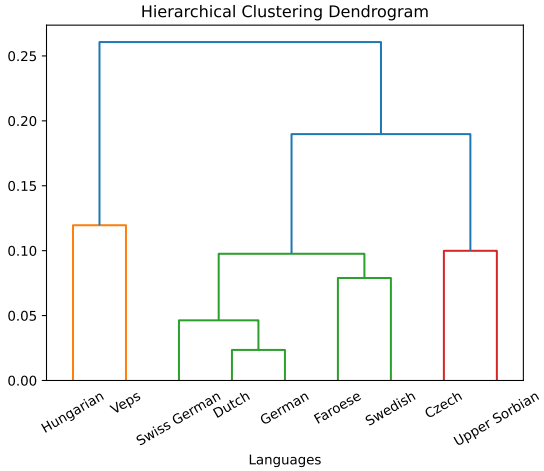


Fig. 1. Agglomerative Hierarchical Clustering dendrogram based on the cosine similarity of the `syntax_knn` vectors of the included languages.

B Appendix

Language	Treebank	Number of tokens	Number of sentences
Czech	CAC	494 142	24 709
Dutch	LassySmall	297 486	17 120
Faroese	OFT	10 002	1 208
German	HDT	3 399 390	189 928
Hungarian	Szeged	42 032	1 800
Swedish	Talbanken	96 859	6 026
Swiss German	ATB	652	98
Upper Sorbian	UFAL	11 196	646
Veps	VWT	1 303	103

Table 1. The specific UD treebanks featured in the project.