It's Basically the Same Language Anyway: the Case for a Nordic Language Model

Magnus Sahlgren[*] RISE Sweden

Fredrik Olsson RISE Sweden

Fredrik Carlsson RISE Sweden

Love Börjeson KB Sweden

Abstract

When is it beneficial for a research community to organize a broader collaborative effort on a topic, and when should we instead promote individual efforts? In this **opinion piece**, we argue that we are at a stage in the development of large-scale language models where a collaborative effort is desirable, despite the fact that the preconditions for making individual contributions have never been better. We consider a number of arguments for collaboratively developing a large-scale Nordic language model, include environmental considerations, cost, data availability, language typology, cultural similarity, and transparency. Our primary goal is to **raise awareness** and **foster a discussion** about our potential impact and responsibility as NLP community.

1 Introduction

Deep Transformer language models have become the weapon of choice in modern NLP (and in AI more generally). There is a rich, and evergrowing, flora of models available, including BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), Electra (Clark et al., 2020), T5 (Raffel et al., 2020), and GPT (2 and 3) (Radford et al., 2019; Brown et al., 2020). These models present slight variations of architectural choices, training objectives, parameter settings, and size and composition of the training data. Despite some internal variation in performance, Transformer language models in general hold state of the art results in basically all NLP benchmarks and evaluation frameworks at the moment (Wang et al., 2018, 2019; Nie et al., 2020).

The downside to this recent development is the computational cost of training deep Transformer models. Starting with BERT-Base with its (now viewed as modest, but at the time of publication seen as substantial) 110 million parameters, and BERT-Large with its 340 million parameters, there has been a virtual explosion in the number of parameters, culminating in the recent GPT-3 with its 175 billion parameters (Brown et al., 2020), GShard with 600 billion parameters (Lepikhin et al., 2021), and the most recent Switch Transformer with a whopping 1,3 *trillion* parameters (Fedus et al., 2021). This development translates into an acute need for access to powerful processing platforms, huge amounts of training data, and an ability to harbor extremely long training times. Taken together, this is a perfect recipe for extreme energy consumption and cost, which risks leading to reduced inclusivity in research on large-scale language models.

There is a budding debate on the environmental and cultural impact of training and using large-scale language models. Two recent examples are Strubell et al. (2019) and Bender et al. (2021); the former analyze the energy consumption and cost of training deep Transformer language models, and the latter voice concerns regarding both the environmental and cultural impact of training and using large-scale language models. We hope to contribute to this discussion by providing a Nordic perspective on the need for large-scale language models. We will assume the position that a collaborative effort towards training a large-scale Nordic language model is something worth striving for. We consider a number of arguments for this position, include environmental considerations, cost, data availability, language typology, cultural similarity, and transparency.

2 Argument 1: The Environment

Strubell et al. (2019) estimate that the CO_2 emission from training a single BERT-Base amounts to roughly 652 kg (1,438 lbs), which is comparable

[*]Corresponding author: magnus.sahlgren@ri.se

to a flight between San Fransisco and New York, or the average emissions resulting from electricity and heating for one person for one year in Stockholm.[1] This is something of a best-case scenario; the authors also calculate that training a BERT-Large with neural architecture search emits something like 284 *tonnes* of CO_2, which is roughly equivalent to the emissions of 56 average persons, throughout a year. An interesting question thus becomes: how much CO_2 emission has been produced as a result of the current development in NLP?

It is of course impossible to get an accurate count on this, but one way to approximate an answer might be to consider how many models have been trained in the world so far. We obviously cannot know this either, but we might be able to get an idea by looking at the number of models published in open source libraries. Luckily, much of the recent development is centered around one such library: the Transformers library of the company Hugging Face.[2] The Transformers library contains (at the time of submission) more than 6,800 models covering a total of 250 languages. A survey carried out by Benaich and Hogarth in the fall of 2020 claims that more than 1,000 companies are using the Transformers library in production, and that it has been installed more than 5 million times.[3]

6,800 models times a low estimate of 652 kg of CO_2 sums to 4,434 tonnes of CO_2 emissions. This is of course an extremely unreliable estimate. Many of the models uploaded to the Transformers library are merely finetuned and not trained from scratch (we have not been able to quantify this proportion). On the other hand, many of the uploaded models are significantly larger than BERT-Base, and one can assume that only a fraction of models that are built are actually uploaded to the Transformers model repository. By comparison, the average Swedish citizen emits around 8 tonnes of CO_2 per year,[4] while RISE (the Research Institutes of Sweden) with approximately 2,800 employees emitted a total of 1,287 tonnes CO_2 during 2019 according to the 2019 annual report.

Counting only the Nordic models uploaded to Hugging Face, there are (at the time of submission) a total of 354 models for the Nordic languages (see Table 1). Based on the assumptions in Strubell et al. (2019), this amounts to more than 230 tonnes of CO_2. By comparison, Anthony et al. (2020) estimates (using slightly different assumptions than Strubell et al. (2019)) that training GPT-3 resulted in at least 85 tonnes of CO_2 emission. Although these estimates are not directly comparable, they indicate that a focused effort to produce a large-scale Nordic language model may lead to a smaller carbon footprint than the current development where we see a steady increase in the number of monolingual models.

Language	Number of models
Swedish	215
Danish	43
Norwegian	33
Icelandic	28
Norwegian Bokmål	12
Norwegian Nynorsk	12
Faroese	11

Table 1: Number of language models available for the Nordic languages via Hugging Face's Transformers library (at the time of submission).

3 Argument 2: Cost

It is anything but cheap to train large-scale language models. The cost for performing a single training pass for the largest T5 model is estimated to be $\$1,3$ million (Sharir et al., 2020), while training GPT-3 is estimated at around $\$4,6$ million.[5] To put these numbers into perspective, the average project funding in the EU Horizon 2020 program is estimated to be around $\$2,1$ million,[6] while the average national research project is typically not more than around $\$150$ *thousand*.[7] This means that, unless you happen to be in the possession of a sizeable computing infrastructure, training models on this scale will be out of the question for most researchers.

However, even with access to suitable GPUs, it is not obvious that it will be possible to train a model on the required scale. Li (2020) estimates

[1] www.regionfakta.com

[2] https://github.com/huggingface/transformers

[3] https://www.stateof.ai/ (slide 127)

[4] www.naturvardsverket.se

[5] lambdalabs.com/blog/demystifying-gpt-3/

[6] accelopment.com/blog/lessons-learnt-from-horizon-2020-for-its-final-2-years/

[7] vr.se/soka-finansiering/beslut/2020-09-08-humaniora-och-samhallsvetenskap.html

that performing a single training run with the full GPT-3 using an NVIDIA Tesla V100 GPU at its theoretical max speed would require 355 years. Assuming access to an NVIDIA DGX-1, which features 8 V100 GPUs, we would still need 44 years to build a replica of GPT-3. The cost of buying a DGX-1 machine is around $129 thousand – i.e. roughly the size of an average national research project.

The sizeable cost (monetary as well as temporal) required to build a large-scale language model effectively excludes a large proportion of the NLP community from training models. This may not be entirely negative, considering the environmental concerns raised in the previous section, but it would be desirable if the production of large-scale language models was more inclusive and collaborative, with transparency and the possibility to influence the procedure even by smaller research groups. A communal effort would not only enable more researchers to have an influence on the model design, but it may also lead to broader usage of the resulting model, thereby reducing the need to constantly build new small (and probably not very useful) models.

4 Argument 3: Data Size and Transfer

It is a known fact that bigger training data leads to improved performance when using statistical learning methods in NLP (Banko and Brill, 2001; Sahlgren and Lenci, 2016). This has been eminently well demonstrated in the context of language models by the recent improvements using models that have been trained on very large data samples (Raffel et al., 2020; Brown et al., 2020). It is a fascinating question whether there *at all* exists sufficiently large text data to build native models for all Nordic languages.

Considering the biggest Nordic language Swedish as an example, Sweden has legal deposit laws installed in 1661 for everything printed. During the the twentieth century it was gradually extended to include sound, moving images and computer games and electronic material. The law for legal deposit of electronic material was added in 2012. As a result, the National Library of Sweden (KB), has vast and ever growing collections, closing in on 26 Petabyte of data.

Though only a fraction of the collections are digitized, the digital collections are nonetheless substantial. KB, through its data lab (KBLab), works continuously to assembly corpora of Swedish texts and to make them available for modeling. The latest corpus of cleaned, edited, raw Swedish text is just over 104 GB of size (corresponding to approximately 1,4 billion sentences and 18,2 billion words). The sources for this corpus are: Swedish Wikipedia 2 GB; Governmental texts 5 GB; Electronic publications 0,4 GB; Social media 5GB; Monographs 2GB, and; Newspapers 90 GB. The corpus currently under construction increases primarily the share of born digital text from legal electronic deposits and is expected to be around 1 TB of cleaned, edited, raw Swedish text (thus approximately 14 billion sentences and 182 billion words). The upper limit (in terms of size) for subsequent corpora is expected to be between $2-5$ TB, depending on the possibilities to transcribe spoken Swedish present in the KB collections.

The situations in the other Nordic countries are similar, relative to the size of the population in the respective countries. There are consequently extensive Danish and Norwegian collections available, whereas the text/data resources in Iceland and Faroe Islands are expected to be substantially smaller. Combining *all* Nordic text resources would likely lead to a fairly substantial data source, likely on the order of Terabytes.

The data conditions for the larger Nordic languages look promising even when considered individually, but it is not obvious that there even exists enough data to train native large-scale models in the smaller Nordic languages. Fortunately, it has been demonstrated that multilingual models improve the performance for languages with less available training data, due to transfer effects (Conneau et al., 2020). In particular, the transfer effects seems to be specifically beneficial for typologically similar languages (Karthikeyan et al., 2020; Lauscher et al., 2020). It is thus likely that in particular Icelandic and Faroese would benefit from a joint Nordic language model.

5 Argument 4: Typology

The Nordic languages belong to one of three Germanic language groups, also referred to as North Germanic languages (in addition to West and now extinct East Germanic). The North Germanic language group is further divided into two branches: East Scandinavian languages, which includes Swedish and Danish, and West Scandina-

vian languages, which contains Norwegian, Icelandic and Faroese. This genealogical categorization is sometimes contrasted with a distinction based on mutual intelligibility, which separates Continental Scandinavian (Swedish, Norwegian and Danish) from Insular Scandinavian (Icelandic and Faroese).

The Nordic languages are so similar from a typological perspective that the language boundaries have been, if not in dispute, at least subject to some discussion (Stampe Sletten et al., 2005). The difference between dialects *within* the Nordic languages is in some cases probably larger than the difference *between* the languages. A telling example is the difference between Norwegian Bokmål, which is very similar to Danish and as such is categorized as an East Scandinavian language, and Nynorsk, which is categorized as a West Scandinavian language. Another example is the difference between Jamtlandish (or Jamska, a dialect spoken in the Swedish region Jämtland, which is categorized as a West Scandinavian language) and standard Swedish (which is East Scandinavian).

From a typological perspective, it thus makes sense to entertain the idea of a joint North Germanic language model, in particular when considering the potential for transfer effects to the smaller Nordic languages. Of course, one can always ask whether we should not aim for a combined Germanic model instead? There will probably be something like an order of magnitude more data available if we consider *all* Germanic languages rather than just the Nordic ones. However, one can expect diminishing returns by adding more data at some point, and it is an interesting (and, as far as we are aware, open) question what is the trade-off between language similarity and data size?

6 Argument 5: Culture

Bender et al. (2021) raise concerns about the considerable anglocentrism of current language models. We agree that this is potentially problematic; most current models are trained on data harvested from the Internet, which we know is produced by certain demographies, and as such is not representative of the general population.[8] A consequence of this is that current language models only encode the perspectives of certain groups of people, and these people tend to *not* belong to marginalized groups. It is well-known that language models encode biases and prejudice that may be problematic (Bordia and Bowman, 2019; May et al., 2019).

Anglocentrism is not necessarily a disqualifying factor for the Nordic countries, some of which (such as Sweden) is sometimes considered to be among the most Americanized countries in the world (Åsard, 2016; Alm, 2003). We generally listen to the same type of music, watch the same type of movies, and watch the same type of TV-shows. We don't, however, have similar political systems (as demonstrated by recent events). By contrast, there is arguably no (significant) difference in culture, politics, or economics between the Nordic countries. In fact, there are probably more cultural differences *within* the countries than between.

A relevant question is how to also include minority languages from other language families, such as Sámi. A natural suggestion for this specific case is to consider a Uralic language model, which would include languages such as Finnish, Hungarian, Estonian, as well as the smaller languages Erzya, Moksha, Mari, Udmurt, Sámi, and Komi.

7 Argument 6: Transparency

The largest concurrent language models are not publicly available. Few have probably missed the controversy surrounding the initial decision of Open AI to *not* release GPT-2 due to concerns of adversarial usages.[9] As we know, GPT-2 was eventually released in full, and there are now GPT-2 models available in many other languages. The original GPT-3 model is however not yet openly available (*Open* AI is beginning to look like a misnomer), but there are several open-source efforts to provide competing, or at least alternative, models.[10,11]

This lack of transparency obviously limits the ability for other researchers not only to investigate this type of model, but also to contribute to its future development. A collaborative Nordic effort would ensure inclusivity in the development, as well as accessibility to the final model.

[8] https://www.pewresearch.org/internet/fact-sheet/social-media/

[9] openai.com/blog/better-language-models/

[10] github.com/EleutherAI/gpt-neo

[11] github.com/sberbank-ai/ru-gpts

8 Conclusions

Based on the considerations raised in this paper, we argue that we – the Nordic NLP community – **should work together to build a truly large-scale Nordic language model**, for the Nordic languages, by Nordic researchers. We believe that such a resource will be extremely beneficial for Nordic NLP, and that it will have the potential to reduce the environmental impact of continuously training new models.

References

Martin Alm. 2003. America and the future of sweden: Americanization as controlled modernization. *American Studies in Scandinavia*, 35(2):64–72.

Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. In *ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems*.

E. Åsard. 2016. *Det blågula stjärnbaneret: USA:s närvaro och inflytande i Sverige*. Carlssons.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, page 26–33, USA. Association for Computational Linguistics.

Emily Bender, Timnit Begru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21*.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. http://arxiv.org/abs/2005.14165 Language models are few-shot learners.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. https://doi.org/10.18653/v1/2020.acl-main.747 Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. http://arxiv.org/abs/2101.03961 Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. {GS}hard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*.

Chuan Li. 2020. Openai's gpt-3 language model: A technical overview. `https://lambdalabs.com/blog/demystifying-gpt-3/`. Accessed: 2021-02-05.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, Open AI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 975–980, Austin, Texas. Association for Computational Linguistics.

Or Sharir, Barak Peleg, and Yoav Shoham. 2020. http://arxiv.org/abs/arXiv:2004.08900 The cost of training nlp models: A concise overview.

Iben Stampe Sletten, Arne Torp, Kaisa Häkkinen, Mikael Svonni, and Carl Christian Olsen. 2005. *Nordens språk - med rötter och fötter*. Number 2004:008 in Nord. Nordisk ministerråd.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3266–3280. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. 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*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.