Should we Stop Training More Monolingual Models, and Simply Use Machine Translation Instead?

Tim Isbister
Peltarion
tim.isbister@peltarion.com

Fredrik Carlsson
RISE
fredrik.carlsson@ri.se

Magnus Sahlgren
RISE
magnus.sahlgren@ri.se

Abstract

Most work in NLP makes the assumption that it is desirable to develop solutions in the native language in question. There is consequently a strong trend towards building native language models even for low-resource languages. This paper questions this development, and explores the idea of simply translating the data into English, thereby enabling the use of pretrained, and large-scale, English language models. We demonstrate empirically that a large English language model coupled with modern machine translation outperforms native language models in most Scandinavian languages. The exception to this is Finnish, which we assume is due to inferior translation quality. Our results suggest that machine translation is a mature technology, which raises a serious counter-argument for training native language models for low-resource languages. This paper therefore strives to make a provocative but important point. As English language models are improving at an unprecedented pace, which in turn improves machine translation, it is from an empirical and environmental stand-point more effective to translate data from low-resource languages into English, than to build language models for such languages.

1 Introduction

Although the Transformer architecture for deep learning was only recently introduced (Vaswani et al., 2017), it has had a profound impact on the development in Natural Language Processing (NLP) during the last couple of years. Starting with the seminal BERT model (Devlin et al., 2019), we have witnessed an unprecedented development of new model variations (Yang et al., 2019; Clark et al., 2020; Raffel et al., 2020; Radford et al., 2019; Brown et al., 2020) with new State Of The Art (SOTA) results being produced in all types of NLP benchmarks (Wang et al., 2018, 2019; Nie et al., 2020).

The leading models are large both with respect to the number of parameters and the size of the training data used to build the model; this correlation between size and performance has been demonstrated by Kaplan et al. (2020). The ongoing scale race has culminated in the 175-billion parameter model GPT-3, which was trained on some 45TB of data summing to around 500 billion tokens (Brown et al., 2020).[1] Turning to the Scandinavian languages, there are no such truly large-scale models available. At the time of writing, there are around 300 Scandinavian models available in the Hugging Face Transformers model repository.[2] Most of these are translation models, but there is already a significant number of monolingual models available in the Scandinavian languages.[3]

However, none of these Scandinavian language models are even close to the currently leading English models in parameter size or training data used. As such, we can expect that their relative performance in comparison with the leading English models is significantly worse. Furthermore, we can expect that the number of monolingual Scandinavian models will continue to grow at an exponential pace during the near future. The question is: do we need all these models? Or even: do we need *any* of these models? Can't we simply translate our data and tasks to English and use some suitable English SOTA model to solve the problem? This paper provides an empirical study of this idea.

[1] The currently largest English model contains 1.6 trillion parameters (Fedus et al., 2021).

[2] huggingface.co/models

[3] At the time of submission, there are 17 monolingual Swedish models available.

Language	Vocab size	Lexical richness	Avg. word length	Avg. sentence length
Swedish	31,478	0.07	4.39	14.75
Norwegian	26,168	0.06	4.21	14.10
Danish	42,358	0.06	4.17	19.55
Finnish	34,729	0.14	5.84	10.69
English	27,610	0.04	3.99	16.87

Table 1: The vocabulary size, Lexical richness, average word length and average sentence length for the Trustpilot sentiment data of each language.

2 Related work

There is already a large, and rapidly growing, literature on the use of multilingual models (Conneau et al., 2020a; Xue et al., 2020), and on the possibility to achieve cross-lingual transfer in multilingual language models (Ruder et al., 2019; Artetxe et al., 2020; Lauscher et al., 2020; Conneau et al., 2020b; Karthikeyan et al., 2020; Nooralahzadeh et al., 2020). From this literature, we know among other things that multilingual models tend to be competitive in comparison with monolingual ones, and that especially languages with smaller amounts of training data available can benefit significantly from transfer effects from related languages with more training data available. This line of study focuses on the possibility to transfer *models* to a new language, and thereby facilitating the application of the model to data in the original language.

By contrast, our interest is to transfer the *data* to another language, thereby enabling the use of SOTA models to solve whatever task we are interested in. We are only aware of one previous study in this direction: Duh et al. (2011) performs cross-lingual machine translation using outdated methods, resulting in the claim that even if perfect translation would be possible, we will still see degradation of performance. In this paper, we use modern machine translation methods, and demonstrate empirically that no degradation of performance is observable when using large SOTA models.

3 Data

In order to be able to use comparable data in the languages under consideration (Swedish, Danish, Norwegian, and Finnish), we contribute a Scandinavian sentiment corpus (ScandiSent),[4] consisting of data downloaded from `trustpilot.com`. For each language, the corresponding subdomain was used to gather reviews with an associated text. This data covers a wide range of topics and are divided into 22 different categories, such as electronics, sports, travel, food, health etc. The reviews are evenly distributed among all categories for each language.

All reviews have a corresponding rating in the range $1-5$. The review ratings were polarised into binary labels, and the reviews which received neutral rating were discarded. Ratings with 4 or 5 thus corresponds to a positive label, and 1 or 2 correspond to a negative label.

To further improve the quality of the data, we apply fastText's language identification model (Joulin et al., 2016) to filter out any reviews containing incorrect language. This results in a balanced set of 10,000 texts for each language, with 7,500 samples for training and 2,500 for testing. Table 1 summarizes statistics for the various datasets of each respective language.

3.1 Translation

For all the Nordic languages we generate a corresponding English dataset by direct Machine Translation, using the Neural Machine Translation (NMT) model provided by Google.[5] To justifiably isolate the effects of modern day machine translation, we restrict the translation to be executed in prior to all experiments. This means that all translation is executed prior to any fine-tuning, and that the translation model is not updated during training.

4 Models

In order to fairly select a representative pre-trained model for each considered Scandinavian language, we opt for the most popular native model according to Hugging Face. For each considered language, this corresponds to a BERT-Base model, hence each language is represented by a Language Model

[4] https://github.com/timpal0l/ScandiSent

[5] https://cloud.google.com/translate/docs/advanced/translating-text-v3

Model name in Hugging Face	Language	Data size
KB/bert-base-swedish-cased	sv	3B tokens
TurkuNLP/bert-base-finnish-cased-v1	fi	3B tokens
ltgoslo/norbert	no	2B tokens
DJSammy/bert-base-danish-uncased_BotXO,ai	da	1.6B tokens
bert-base-cased	en	3.3B tokens
bert-base-cased-large	en	3.3B tokens
xlm-roberta-large	multi	295B tokens

Table 2: Models used in the experiments and the size of their corresponding training data. 'B' is short for billion.

Model	sv	no	da	fi	en
BERT-sv	96.76	89.32	90.68	83.40	86.76
BERT-no	90.40	95.00	92.52	83.16	78.52
BERT-da	86.24	89.16	94.72	80.16	85.28
BERT-fi	90.24	86.36	87.72	**95.72**	84.32
BERT-en	85.72	87.60	87.72	84.16	96.08
BERT-en-Large	91.16	91.88	92.40	89.56	**97.00**
Translated Into English					
BERT-sv	88.24	87.80	89.68	83.60	-
BERT-no	88.40	86.80	88.44	80.72	-
BERT-da	88.24	84.20	89.12	83.32	-
BERT-fi	90.04	90.08	89.36	86.04	-
BERT-en	95.76	95.48	95.96	92.96	-
BERT-en-Large	**97.16**	**96.56**	**97.48**	94.84	-

Table 3: Accuracy for monolingual models for the native sentiment data (upper part) and machine translated data (lower part). Underlined results are the best results per language in using the native data, while boldface marks the best results considering both native and machine translated data.

Model	sv	no	da	fi	en
XLM-R-large	**97.48**	**97.16**	97.68	95.60	**97.76**
Translated Into English					
XLM-R-large	97.04	96.84	**98.24**	95.48	-

Table 4: Accuracy on the various sentiment datasets using XLM-R-Large

of identical architecture. The difference between these models is therefore mainly in the quantity and type of texts used during training, in addition to potential differences in training hyperparameters.

We compare these Scandinavian models against the English BERT-Base and BERT-Large models by Google. English BERT-Base is thus identical in architecture to the Scandinavian models, while BERT-Large is twice as deep and contains more than three times the amount of parameters as BERT-Base. Finally, we include XLM-R-Large, in order to compare with a model trained on significantly larger (and multilingual) training corpora.

Table 2 lists both the Scandinavian and English models, together with the size of each models corresponding training corpus.

5 Experiments

5.1 Setup

We fine-tune and evaluate each model towards each of the different sentiment datasets, using the hyperparameters listed in Appendix 5. From this we report the binary accuracy, with the results for the BERT models available in Table 3, and the XLM-R results in Table 4.

5.2 Monolingual Results

The upper part of Table 3 shows the results using the original monolingual data. From this we note a clear diagonal (marked by underline), where the native models perform best in their own respective language. Bert-Large significantly outperforms BERT-Base for all non-English datasets, and it also performs slightly better on the original English data.

Comparing these results with the amount of training data for each model (Table 1), we see a correlation between performance and amount of pre-training data. The Swedish, Finnish and English models have been trained on the most amount of data, leading to slightly higher performance in their native languages. The Danish model which has been trained on the least amount of data, performs the worst on its own native language.

For the cross-lingual evaluation, BERT-Large clearly outperforms all other non-native models. The Swedish model reaches higher performance on Norwegian and Finnish compared to the other non-native Scandinavian models. However, the Norwegian model performs best of the non-native models on the Danish data. Finally, we observe an interesting anomaly in the results on the English data, where the Norwegian model performs considerably worse than the other Scandinavian models.

5.3 Translation Results

The results for the machine translated data, available as the lower part of Table 3, show that BERT-Large outperforms all native models on their native data, with the exception of Finish. The English BERT-Base reaches higher performance on the machine translated data than the Norwegian and Danish models on their respective native data. The difference between English BERT-Base using the machine translated data, and the Swedish BERT using native data is about 1% unit.

As expected, all Scandinavian models perform significantly worse on their respective machine translated data. We find no clear trend among the Scandinavian models when evaluated on translated data from other languages. But we note that the Danish model performs better on the machine translated Swedish data than on the original Swedish data, and the Finnish model also improves its performance on the other translated data sets (except for Swedish). All models (except, of course, the Finnish model) perform better on the machine translated Finnish data.

Finally, 4 shows the results from XLM-R-Large, which has been trained on data several orders of magnitude larger than the other models. XLM-R-Large achieves top scores on the sentiment data for all languages except for Finnish. We note that XLM-R produces slightly better results on the native data for Swedish, Norwegian and Finnish, while the best result for Danish is produced on the machine translated data.

6 Discussion & Conclusion

Our experiments demonstrate that it is possible to reach better performance in a sentiment analysis task by translating the data into English and using a large pre-trained English language model, compared to using data in the original language and a smaller native language model. Whether this result holds for other tasks as well remains to be shown, but we see no theoretical reasons for why it would not hold. We also find a strong correlation between the quantity of pre-training data and downstream performance. We note that XLM-R in particular performs well, which may be due to data size, and potentially the ability of the model to take advantage of transfer effects between languages.

An interesting exception in our results is the Finnish data, which is the only task for which the native model performs best, despite XLM-R reportedly having been trained on more Finnish data than the native Finnish BERT model (Conneau et al., 2020a). One hypothesis for this behavior can be that the alleged transfer effects in XLM-R hold primarily for typologically similar languages, and that the performance on typologically unique languages, such as Finnish, may actually be negatively affected by the transfer. The relatively bad performance of BERT-Large on the translated Finnish data is likely due to insufficient quality of the machine translation.

The proposed approach is thus obviously dependent on the existence of a high-quality machine translation solution. The Scandinavian languages are typologically very similar both to each other and to English, which probably explains the good performance of the proposed approach even when using a generic translation API. For other languages, such as Finnish in our case, one would probably need to be more careful in selecting a suitable translation model. Whether the suggested methodology will be applicable to other language

pairs thus depends on the quality of the translations and on the availability of large-scale language models in the target language.

Our results can be seen as evidence for the maturity of machine translation. Even using a generic translation API, we can leverage the existence of large-scale English language models to improve the performance in comparison with building a solution in the native language. This raises a serious counter-argument for the habitual practice in applied NLP to develop native solutions to practical problems. Hence, we conclude with the somewhat provocative claim that it might be unnecessary from an empirical standpoint to train models in languages where:

1. there exists high-quality machine translation models to English,

2. there does not exist as much training data to build a language model.

In such cases, we may be better off relying on existing large-scale English models. This is a clear case for practical applications, where it would be beneficial to only host one large English model and translate all various incoming requests from different languages.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. 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. 2020a. 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.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, 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.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, page 429–433, USA. Association for Computational Linguistics.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. http://arxiv.org/abs/2001.08361 Scaling laws for neural language models.

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.

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.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-Shot Cross-Lingual Transfer with Meta Learning. In *Proceedings of EMNLP*. 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.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. http://arxiv.org/abs/2010.11934 mT5: A massively multilingual pre-trained text-to-text transformer. ArXiv:2010.11934.

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.

A Training Details

Parameters	Value
train_epochs	2
early_stopping	false
optimizer	AdamW
learning_rate	4e-5
batch_size	512
max_seq_length	128
max_grad_norm	1.0

Table 5: Training hyperparameters for the sentiment classification experiments.