Are There Any Limits to English-Swedish Language Transfer? A Fine-grained Analysis Using Natural Language Inference

Felix Morger
Språkbanken, Department of Swedish, Multilingualism, Language Technology
Gothenburg university
felix.morger@gu.se

Abstract

The developments of deep learning in natural language processing (NLP) in recent years have resulted in an unprecedented amount of computational power and data required to train state-of-the-art NLP models. This makes lower-resource languages, such as Swedish, increasingly more reliant on language transfer effects from English since they do not have enough data to train separate monolingual models. In this study, we investigate whether there is any potential loss in English-Swedish language transfer by evaluating two types of language transfer on the GLUE/SweDiagnostics datasets and comparing between different linguistic phenomena. The results show that for an approach using machine translation for training there is no considerable loss in overall performance nor by any particular linguistic phenomena, while relying on pretraining of a multilingual model results in considerable loss in performance. This raises questions about the role of machine translation and the use of natural language inference (NLI) as well as parallel corpora for measuring English-Swedish language transfer.

1 Introduction

The leveraging of *knowledge transfer* has been a pivotal development in deep learning and in natural language processing (NLP). The core idea is that the knowledge of one source domain can be transferred to another target domain. In the age of deep learning, this usually means training models on large quantities of data on some generic or widely-applied task, such as language modeling or machine translation, and fine-tuning the model on new data or another downstream task. This does not only lead to an immediate performance boost on the new dataset or downstream task, but also saves computational resources of the user as well as simplifies the implementation and problem solving procedure.

A crucial type of knowledge transfer in NLP is that of *language transfer*, which refers to the leveraging of high-resource languages, most notably English, to solve tasks for lower-resource languages. It has been studied in multiple different architectures and applications of NLP, such as in word embedding architectures and machine translation. The most used type of language transfer, however, is arguably that of pretrained language models, which generic pre-training allows for fine-tuning a wide range of down-stream tasks (both in the areas of text generation and text classification).

As deep learning has scaled up in recent years as a result of more computationally efficient hardware and model architectures, most notably with the release of the Transformer (Vaswani et al., 2017), so has the data needed for training. For lower-resource languages these developments come in the shape of a double-edged sword: On the one hand, the breadth of data and contextual information encoded in these models enables a high level of knowledge transfer, on the other hand, they become even more reliant on high-resource languages like English to use the most recent state-of-the-art language models. This leaves lower-resource languages the options of either opting for smaller models and, thus, missing out on the latest progress or to rely heavily on language transfer. Seeing that the latter option seems most plausible, this raises the question to what limits there are to language transfer from a linguistic perspective, but also in terms of potential dangers of political, gender and cultural biases (Bender et al., 2021).

In this study, we focus on English-Swedish language transfer. Swedish, a mid-resource

Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023), pages 30–41, May 22, 2023 ©2023 Association for Computational Linguistics

language, is an interesting example of a language becoming more reliant on language transfer. This is because Swedish has had enough available pre-training data to train monolingual language models, like BERT (Devlin et al., 2019; Malmsten et al., 2020), but not enough for the most recent GPT-3 model (Brown et al., 2020). Also, a significant amount of down-stream tasks in English are not trainable in a supervisable way in Swedish because corresponding annotated data do not exist for the language. In order to test the potential limits of English-Swedish language transfer, we use the GLUE/SweDiagnostics parallel corpora to compare two different types of language transfer. In short, we aim to answer the following research questions:

Q1: Is there a loss in performance in English-Swedish language transfer in the context of natural language inference (NLI)?
Q2: Which type of language transfer works best (machine translation or multilingual pre-training)?
Q3: Are there linguistic phenomena for which language transfer works less effectively?

With this work, we aim to shed light on the English-Swedish language transfer capabilities of pretrained language models and, thus, provide direction for applying and evaluating language transfer in the future.

2 Related Work

This work focuses on the English-Swedish language transfer of pretrained language models in which NLI is used as the measurement. As such, this section will focus on work relating to language transfer and NLI.

2.1 Language transfer

Transfer learning is an attractive solution to the inherent problem of the lack of language data (raw and annotated) for lower-resource languages. For this reason, a multitude of techniques in different areas of research have been proposed to leverage higher-resource languages for lower-resource languages. These range from creating bilingual dictionaries, which can be created unsupervised from just monolingual embedding spaces even for distant language pairs like English-Chinese Lample et al. (2018), to using back-translation between languages to increase performance on machine translation (Sennrich et al., 2016).

One of the most known types of transfer learning is that of Transformer-based multilingual models, such as multilingual BERT (Devlin et al., 2019) or Google Neural Machine Translation (GNMT) (Wu et al., 2016), which train the same models on multiple languages. These models have shown to have transfer effects between the languages in the model. For example, Pires et al. (2019) have shown that fine-tuning XLM (Conneau and Lample, 2019) on natural entity recognition and part-of-speech tagging have transfer effects between closely related languages. This has also been confirmed for multilingual BERT in the context of NLI.[1] Given these results, we expect to see a high degree of transfer between English and Swedish in this study.

The role of language transfer for Swedish has become increasingly relevant for Swedish NLP in recent years. Firstly, the question has been raised of which languages (Swedish, English or other Nordic languages) to train new models on (Sahlgren et al., 2021) as creating pre-trained language models comes with a large financial and environmental cost as well as the need for a lot of data. Secondly, there is an open question to the extent of which machine translation could be deployed for Swedish NLP as Swedish-English machine translation of input and output data has been shown to be effective for Swedish sentiment analysis (Isbister et al., 2021). Thirdly, there is an immediate question of how to leverage language transfer from English seeing that the newly released SuperLim, a standardized benchmark for Swedish (Adesam et al., 2020), has many datasets with little or no training data at all. With this study, we contribute to answering these questions. Specifically, we also test if the findings of Isbister et al. (2021) hold for English-Swedish language transfer in the context of NLI, an arguably higher-level reasoning task than sentiment analysis.

2.2 Natural language inference

Developing datasets for natural language inference (NLI), also called textual entailment, has been a natural endeavor in the NLP community based on the assumption that the identification and resolving of latent logical relations are necessary for language processing. Although, NLI it-

[1] https://github.com/google-research/bert/blob/master/multilingual.md#results (accessed 2022-12-01).

self is usually not a practical task to solve on an application-level, it measures the semantic inference needed to solve other tasks such as question-answering, reading comprehension and sentiment analysis.

For English, many datasets have been developed for NLI. An early dataset is FraCaS (Cooper et al., 1996), which similarly to the later GLUE diagnostics (Wang et al., 2018), targets different categories of logical relations, such as comparatives and quantifiers. In recent years, with the development of deep learning and the need for big data, much larger datasets for NLI have been developed, most notably the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015) and the Multi-Genre Natural Language Inference (MultiNLI or MNLI) dataset (Williams et al., 2018), which use crowdworkers to curate a large amount of sentence pairs (see Section 3.1 for more details).

While NLI has been viewed as an important task for measuring natural language understanding (NLU), as illustrated by its inclusion in the GLUE and SuperGLUE benchmarks (Wang et al., 2018, 2019), NLI datasets have been extensively scrutinized. Multiple studies have shown that performance on these tasks can remain high even after significant transformations of the input, indicating that certain textual artifacts can be used to solve the task rather than logical reasoning. For example, high performance can still be achieved when words of specific wordclasses are dropped (Talman et al., 2021), words in a sentences are shuffled randomly (Pham et al., 2021) or when only the hypothesis is used to predict entailment (Gururangan et al., 2018). At the same time, however, NLI datasets do not transfer well out-of-domain to other NLI datasets, as shown by Talman and Chatzikyriakidis (2019). The GLUE Diagnostic dataset, which is used in this study, has, however, shown to be more robust against such data artifacts, at least when tested against the hypothesis-only baseline of Gururangan et al. (2018) (Wang et al., 2018).

Challenges, thus, remain in developing NLI datasets for measuring logical inference. In this study, we add a different angle to this question by looking at how sensitive NLI datasets can be to English-Swedish machine translation.

3 Datasets & models

For the following study, we use a collection of different NLI datasets and pretrained language models to assess language transfer capabilities from English to Swedish. Table 1 lists the datasets and Table 2 lists the models used in this study.

We use the datasets MNLI and SNLI for training, which due to their unmatched size are most suitable for fine-tuning large pretrained language models. For testing, we use the GLUE/SweDiagnostic parallel corpora, which we use to both evaluate performance as well as to make fine-grained analysis of specific language phenomena.

When it comes to choosing models, we select them based on the criteria that their architectures are 1) available in both languages and 2) directly comparable in terms of architectural size. For these purposes, the `BERT`-model is naturally favored since its `base`-version is available in both Swedish and English (see Section 3.2).

The following subsections give a closer overview of the datasets (Section 3.1) and models (Section 3.2).

3.1 Datasets

SNLI

The SNLI Corpus (Bowman et al., 2015) is the largest NLI corpus to date. It consists of 570K human-written sentence pairs labeled either as contradiction, neutral or entailment. It was curated through crowd-workers (Amazon Mechanical Turk), where each participant was asked to make an entailment, neutral and contradiction hypothesis from scene descriptions in the Flickr30K corpus, resulting in a completely balanced dataset. Additionally, 10% of the data were validated by four more annotator crowd-workers.

MNLI

The MNLI corpus (Williams et al., 2018) is a collection of 433K sentence pairs. It was produced using similar methodology as SNLI: Using crowd-workers to create hypotheses from a premise and validating 10% of the resulting labels using other crowd-workers. What differs from SNLI, however, is that MNLI draws its premises from ten different text genres ranging from transcribed telephone calls to magazine articles, and without accompanying images. Five of these genres, however, are a separate *mismatched* subset of the data

Dataset	Available language(s)	Split(s)	Split size(s)
MultiNLI matched	English	train, dev, test (hidden)	392,702 / 10,000 / 9796
MultiNLI mismatched	English	dev, test (hidden)	10,00 / 9,847
SNLI	English	train, dev, test	550,153 / 10,000 / 10,000
GLUE/SweDiagnostics	English, Swedish	test	1,104

Table 1: Datasets used in this study. Splits refers to the available splits of the model, and split sizes are the number of samples in each given split.

and only come in the development and test split in order to test a system's performance on out-of-domain data.

GLUE/SweDiagnostics

The GLUE diagnostic is an NLI dataset, which is included in the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) evaluation benchmarks. The dataset consists of 1106 hand-picked sentence pairs from four different text sources (ACL proceedings, Artificial, News, Reddit and Wikipedia). Each sample is labeled with additional linguistic phenomena, which the inference relation relies on. The hypothesis and premise only differ in the targeted linguistic phenomenon, so if the inference is not correctly classified, the assumption is that the system cannot handle the targeted linguistic phenomenon. There are linguistic phenomena of 33 different fine-grained categories across four coarse-grained categories (lexical semantics, predicate-argument structure, logic and common sense). Table 4 in the Appendix shows the full list of fine-grained categories as well as their frequency in the dataset.

Because of this human-curated process, the dataset does not represent a natural language distribution (Belinkov and Glass, 2019, Section 4) from the corpora it draws its samples from and is limited in testing for overall performance. Rather, it *diagnoses* the system's ability to solve specific language phenomena. Since the dataset is unbalanced, the metric used in GLUE/SuperGLUE is R_3 (Gorodkin, 2004) a three-class generalization of Matthews correlation coefficient (MCC), which we also use in this study.

SweDiagnostics is the Swedish, manually translated version of GLUE diagnostics. It is part of the SuperLim project (Adesam et al., 2020), which is an evaluation benchmark for Swedish.

3.2 Models

In this study, we use three BERT models to compare for transferability: The Swedish monolingual BERT-base model of the National Library of Sweden (Malmsten et al., 2020), the original monolingual English BERT-base model of Devlin et al. (2019) and the multilingual BERT model. The reasons for choosing these models is because they (1) are architecturally of the same size (12 layer, 768-hidden, 12-heads, 110M parameters) and (2) use roughly the same source of data for training. See Table 2 for comparison.

Model name	Trained on	Data sources
KB-BERT	Swedish	Books, news, government publications, Swedish Wikipedia, internet forums
mBERT	100 languages	Wikipedia of 100 languages
BERT	English	BooksCorpus, English Wikipedia

Table 2: Models used for this study. We use *Model name* to refer to specific models in the paper.

4 Method

Our methodology consists of firstly training different BERT models using two types of language transfer (Section 4.1) and, secondly, evaluating their performance on the GLUE/SweDiagnostics parallel corpora (Section 4.1).[2] Table 3 illustrates the training and evaluation setup in detail.

4.1 Training procedure

Using the datasets and models laid out in Section 3.1 and 3.2, we compare two types of English-Swedish language transfer: multilingual pre-training and fine-tuning on English-Swedish machine translated data. For this, we deploy three different training procedures: The first is to

[2] The code for the experiments is available online: https://github.com/felixhultin/nli-lang-transfer-experiments

fine-tune the multilingual BERT model on English and to rely only on the language transfer effects of pre-training. The second is to fine-tune the Swedish monolingual BERT model on English-Swedish machine translated data, which has been shown to be efficient for Swedish-English language transfer in the context of sentiment analysis (Isbister et al., 2021). The third is to also fine-tune the multilingual BERT on English-Swedish machine translated data for complete comparison. For machine translation, we use the OPUS-MT framework (Tiedemann and Thottingal, 2020). Finally, as a baseline, we compare the results to the performance on the monolingual English BERT model.

For fine-tuning, we use the same training regimen for all models. We use Adam (Kingma and Ba, 2014) with an intial learning rate of 10^{-5}, a dropout (Srivastava et al., 2014) probability of 0.1, a batch size of 80 and train for 3 epochs.

The resulting models are, henceforth, referred to by the pretrained language model name (KB-BERT/mBERT/BERT) and the dataset (mnli or snli) it has been fine-tuned on, as specified in Table 3. For example, KB-BERT, fine-tuned on English-Swedish machine translated SNLI is called KB-BERT.snli-sv and BERT fine-tuned on the English MNLI is called BERT.mnli

4.2 Evaluation

We evaluate on the GLUE/SweDiagnostic dataset by overall performance and fine-grained categories in order to see to which degree specific linguistic phenomena transfer from English to Swedish. Since the distribution of labels (i.e. entailment, neutral and contradiction) is unbalanced, we use R_3 (Gorodkin, 2004), the three-class generalization of the Matthews correlation coefficient. As a sanity check, we evaluate all models on both the GLUE- and SweDiagnostics dataset.

For reference, we also provide the results of the evaluation on the MNLI and SNLI datasets in Figure 4 in the Appendix. However, since the Swedish test data here is machine translated into Swedish and its translation quality has not been manually checked, it is not known which inference relations still hold after translation. Therefore, it is impossible to know to which extent the results reflects machine translation quality or the model's ability to solve the task and should be taken with a grain of salt.

5 Results

5.1 Overall performance

Figure 1 shows the results on the GLUE/SweDiagnostic dataset of all the models. As expected, the English monolingual BERT baseline model does better on GLUE Diagnostics than on SweDiagnostics (+0.18 MCC on BERT.mnli and +0.19 on BERT.snli), while the Swedish monolingual KB-BERT model does better on SweDiagnostics than on GLUE Diagnostics (+0.18 MCC on KB-BERT.mnli-sv and +0.09 on KB-BERT.snli-sv). The multilingual mBERT model performs more evenly on both GLUE Diagnostics and SweDiagnostics (e.g. 0.05 MCC difference on mBERT.mnli and 0.01 on mBERT.snli)), however, it also does better on GLUE Diagnostics when fine-tuned on original English data (mBERT.mnli/snli) and better on SweDiagnostics when fine-tuned on Swedish machine translated data (mBERT.mnli-sv/snli-sv).

These results show that for both the monolingual Swedish KB-BERT model and the multilingual BERT model, fine-tuning on machine translated data achieves complete level of language transfer — at least in terms of performance (Q1). In fact, we even see the Swedish KB-BERT.mnli-sv model perform slightly better (+0.02 MCC) than the English BERT.mnli baseline. While language transfer only from pre-training gives an immediate and considerable performance boost, it does not reach the same extent of completeness as fine-tuning on machine translated data (Q1 & Q2). For example, compare the +0.4 MCC better performance of mBERT.mnli-sv to mBERT.mnli on SweDiagnostics. Similar trends can be observed when evaluating model performance on MNLI and SNLI (see Figure 4 in the Appendix), which further consolidates the fact that training on English-Swedish machine translated data does not impact existing performance on English.

5.2 Performance by linguistic phenomena

When comparing performance between NLI datasets, we see that the models fine-tuned on MNLI perform considerably better than those fine-tuned on SNLI — indicating that the multi-genre MNLI gives rise to broader generalization. In the light of this, we now focus on these models to compare between fine-grained linguistic phenom-

Model	Training data	Test data
KB-BERT	MNLI (MT en-sv)	GLUE/SweDiagnostics
	SNLI (MT en-sv)	
mBERT	MNLI	
	SNLI	
	MNLI (MT en-sv)	
	SNLI (MT en-sv)	
BERT	MNLI	
	SNLI	

Table 3: Training and evaluation setup for the different language transfer procedures and the English baseline. “MT en-sv” is short for machine translated from English to Swedish.

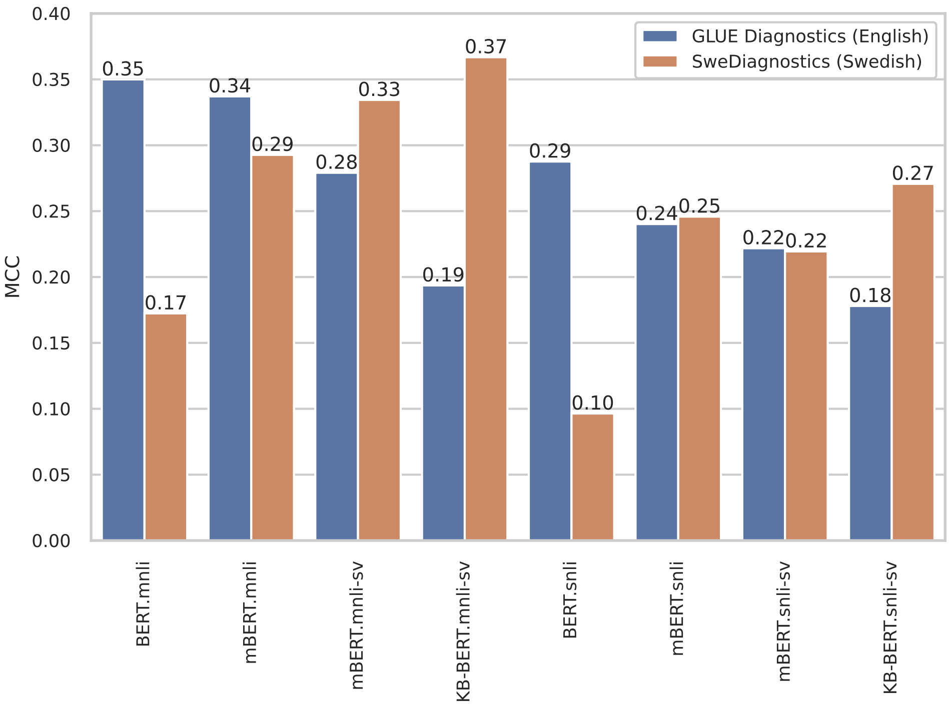
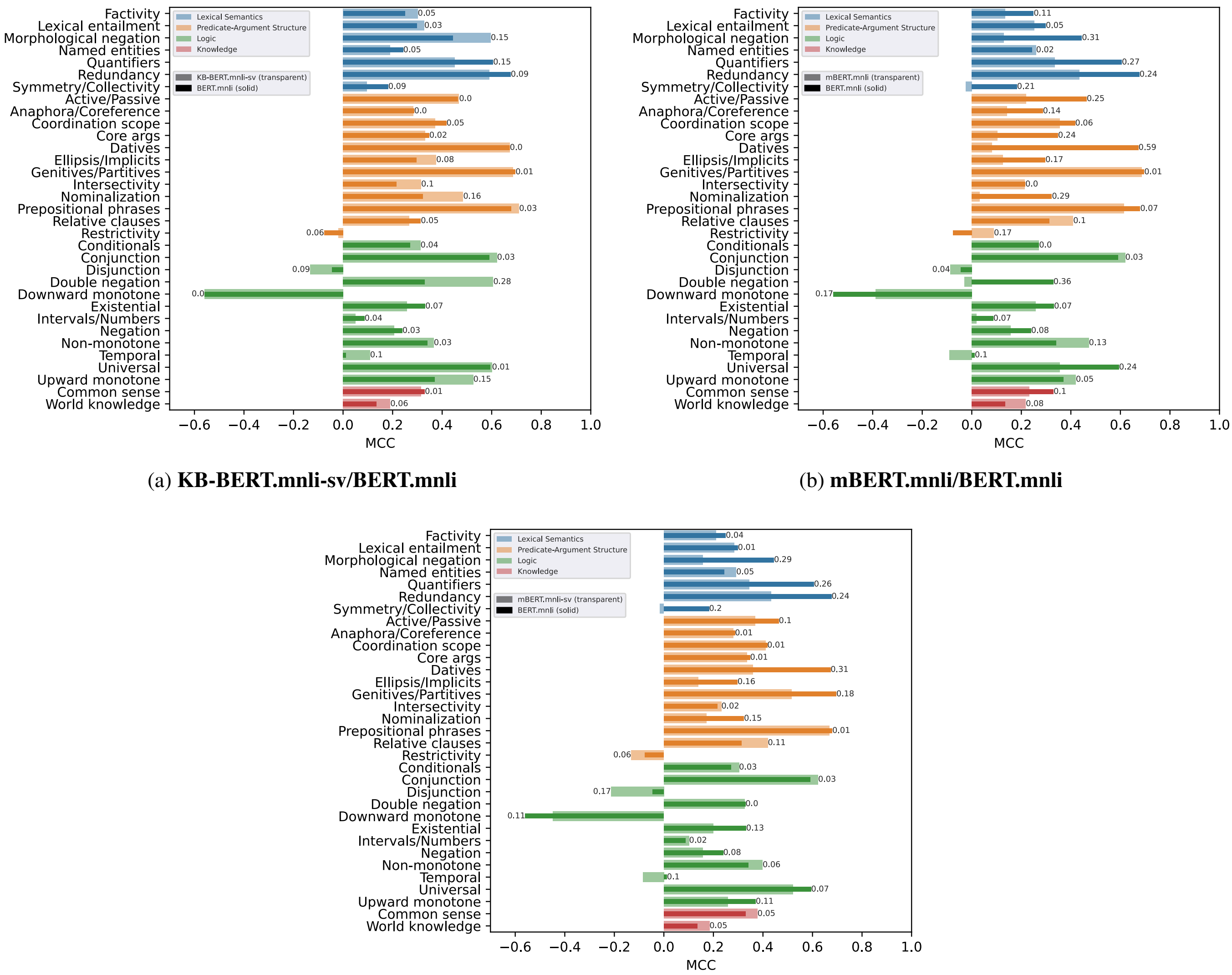


Figure 1: Model performance (MCC) on GLUE/SweDiagnostics (grouped by color).

ena more closely. Figure 2a compares the results of the Swedish `KB-BERT.mnli-sv` model to the baseline `BERT.mnli` model, evaluated on the SweDiagnostics and GLUE Diagnostic dataset, respectively. Here, there is comparable performance *between models* by fine-grained linguistic phenomena despite the fact that there are big differences *between fine-grained categories*. Only five fine-grained phenomena differ by more than 0.1 MCC and only one phenomenon (Double Negation) differ by more than 0.2 MCC.

Figure 2b compares `mBERT.mnli` to `BERT.mnli`, which are also evaluated on the SweDiagnostics and GLUE Diagnostic dataset, respectively. When comparing these results to Figure 2a, we see that the `BERT.mnli` baseline model performs markedly better than the multilingual `mBERT.mnli` model by many linguistic phenomena. Some phenomena stand out, such as “Datives”, “Morphological negation” and “Quantifiers”.

Figure 2c compares `mBERT.mnli-sv` to `BERT.mnli`. The performance here is not on the same level as the `KB-BERT.mnli-sv`, but we see that the gap is narrowed in some categories seen in Figure 2b by training on English-Swedish machine translated data. “Active Passive”, “Core args”, “Datives” and “Double negation” stand out in particular, which perform 0.24, 0.23, 0.28 and 0.36 MCC better than when relying only on pre-



(a) **KB-BERT.mnli-sv/BERT.mnli**

(b) **mBERT.mnli/BERT.mnli**

(c) **mBERT.mnli-sv/BERT.mnli**

Figure 2: Overlapping barchart comparing Matthews correlation coefficient (MCC) of two models by fine-grained phenomenon from the GLUE Diagnostic dataset, The results of one model is in transparent color bars and the other in **thick color bars**. Bars are grouped by coarse-grained phenomena with different colors and the bar labels indicate the absolute difference in MCC between models.

training (see Figure 2b).

Given the high similarity in performance on the GLUE/SweDiagnostic dataset of the English baseline `BERT.mnli` and `KB-BERT.mnli`, we check explicitly if this is because they make the same predictions. We therefore test prediction agreement between the models as well as the gold labels for reference. Figure 3 shows the results and confirms a strong model agreement (0.71 MCC) between `KB-BERT.mnli` and `BERT.mnli` as well as a slightly lower model agreement (0.63 MCC) between `BERT.mnli` and `mBERT.mnli`. Similar transfer effects from machine translation can be seen in `mBERT.mnli-sv`. Thus, this confirms that the similarity in performance of the models is largely because they are making the same predictions.

6 Discussion

The results show that even for a high-level reasoning task such as NLI, English-Swedish language transfer can be made without any considerable loss in performance (Q1). This finding is further solidified when comparing by different linguistic phenomena where the performance is comparable to an equivalent English baseline (Q3). When com-

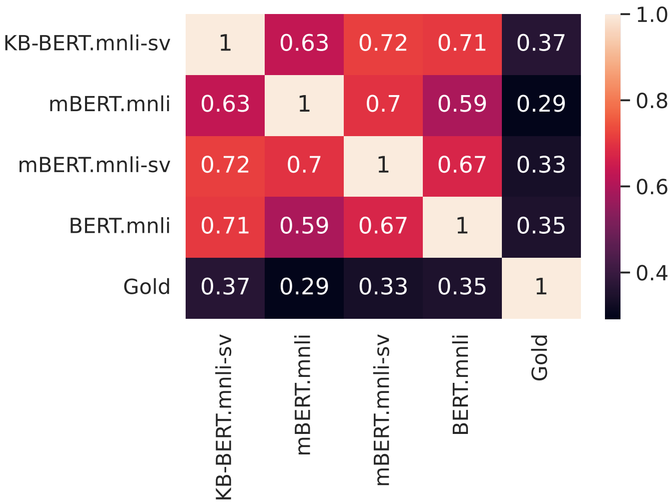


Figure 3: Model prediction agreement (MCC) of the MNLI-based models on the GLUE/SweDiagnostic dataset. "Gold" are the true labels.

paring different types of language transfer, we see that that fine-tuning on English-Swedish machine translated data performs markedly better than relying only on language transfer from multilingual pre-training. Thus, it confirms the findings of Isbister et al. (2021) that leveraging machine translation between English and Swedish is an effective method of language transfer (Q2).

The high similarity in model performance and prediction agreement suggests that the information used by the BERT model to predict inference, is not lost after English-Swedish machine translation. Conversely, this also suggests that whatever language information is lost after machine translation, it is not essential for predicting inference. This further highlights, as previous studies have, the brittleness of NLI as a measurement of natural language understanding. If we were to assume that (a) the model uses logical reasoning to solve NLI and (b) important entailment information is lost after machine translation, then the performance after language transfer should vary considerably. However, since that is not the case, it seems more plausible that the model uses other textual artifacts to predict inference. Alternatively, the entailment relation between the hypothesis and premise does not change considerably after machine translation. Since we do not know, however, the extent to which entailment relations hold after machine translation, we cannot know for certain and, thereby, could be a question for future studies to explore.

Given the observations above, it is also important to take into account that the results of this study are only as generalizeable as the peculiarities of the GLUE/SweDiagnostic dataset. Its relatively small size (1106 sentence pairs), choice of genres, linguistic phenomena and annotation procedure might not generalize to all cases of NLI language transfer. In particular, since SweDiagnostics was manually translated into Swedish from English, it most likely has *translationese* (Gellerstam, 1986) elements in it and might, thus, naturally be biased towards machine translation output. Furthermore, translating premise and hypothesis independently, has been shown to reduces lexical overlap between the sentences (Artetxe et al., 2020), which could help the model not overfit on spurious annotation artefacts. Until a unique NLI dataset for Swedish is created, which samples are taken from naturally occurring spoken or written Swedish, we cannot know the extent to which this impacts the results.

7 Conclusion

In this study, we show that for the high-level reasoning task of NLI, English-Swedish language transfer can be done without any considerable loss in performance. We also see that for a model which uses machine translation for training, there is no considerable loss by any specific linguistic phenomenon. Meanwhile, a multilingual model which only relies on pre-training for language transfer does not see the same level of language transfer.

Given the increasing reliance on English-Swedish language transfer as a result of the development towards larger models with need for more training data in NLP, we see a need for further studies into the potential effects of language transfer on Swedish. In this effort, understanding the role of English-Swedish translation as well as comparing these results to datasets that are based on naturally occurring written or spoken Swedish, will be essential to understand the true impact of English linguistic and cultural influences on English-Swedish language transfer. Finally, applying similar studies to newer and larger pretrained language models, such as GPT-SW3, will become even more important as they will be used more broadly in the future.

References

Yvonne Adesam, Aleksandrs Berdicevskis, and Felix Morger. 2020. Swedishglue–towards a swedish test set for evaluating natural language understanding models.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

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.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95.

Jan Gorodkin. 2004. Comparing two k-category assignments by a k-category correlation coefficient. *Computational biology and chemistry*, 28(5-6):367–374.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Tim Isbister, Fredrik Carlsson, and Magnus Sahlgren. 2021. Should we stop training more monolingual models, and simply use machine translation instead? *arXiv preprint arXiv:2104.10441*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden – making a swedish bert.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. 2021. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1145–1160, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Magnus Sahlgren, Fredrik Carlsson, Fredrik Olsson, and Love Börjeson. 2021. It's basically the same language anyway: the case for a Nordic language model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 367–372, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Aarne Talman, Marianna Apidianaki, Stergios Chatzikyriakidis, and Jörg Tiedemann. 2021. NLI data sanity check: Assessing the effect of data corruption on model performance. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 276–287, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Aarne Talman and Stergios Chatzikyriakidis. 2019. Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

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*, volume 32. 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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Appendix

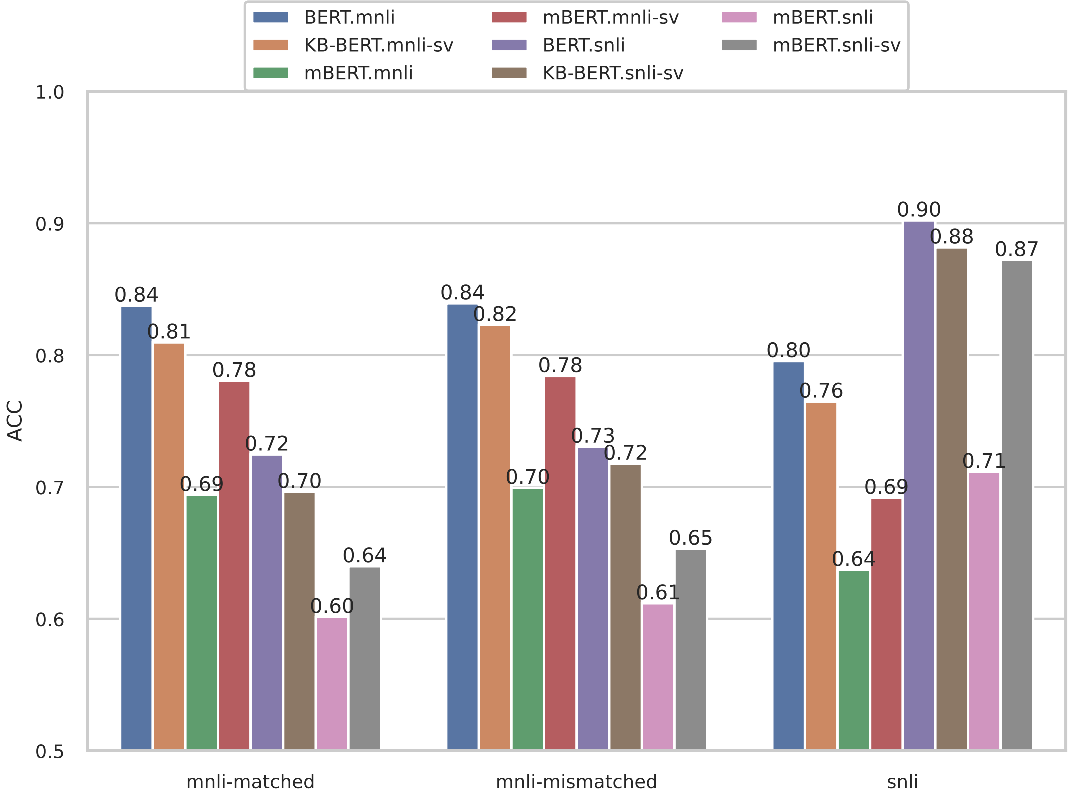


Figure 4: Accuracy on the MNLI and SNLI datasets. Bars are grouped by task and color indicates model. `mBERT`, and `KB-BERT`, are evaluated on the English-Swedish machine translated version of MNLI and SNLI while the English baseline `BERT` is evaluated on the original English version.

Coarse-grained	Fine-grained	Size	Neutral	Entailment	Contradiction
Lexical Semantics	**Factivity**	68	37	17	14
	Lexical entailment	140	37	49	54
	Morphological negation	26	2	14	10
	Named entities	36	12	18	6
	Quantifiers	52	18	14	20
	Redundancy	26	2	24	0
	Symmetry/Collectivity	28	8	20	0
Predicate-Argument Structure	**Active/Passive**	34	17	15	2
	Anaphora/Coreference	58	22	24	12
	Coordination scope	40	16	14	10
	Core args	52	15	27	10
	Datives	20	4	14	2
	Ellipsis/Implicits	34	4	16	14
	Genitives/Partitives	20	2	16	2
	Intersectivity	46	25	19	2
	Nominalization	28	4	18	6
	Prepositional phrases	68	32	34	2
	Relative clauses	32	16	12	4
	Restrictivity	26	9	17	0
Logic	**Conditionals**	32	8	18	6
	Conjunction	40	15	15	10
	Disjunction	38	17	15	6
	Double negation	28	2	22	4
	Downward monotone	30	17	13	0
	Existential	20	9	7	4
	Intervals/Numbers	38	11	9	18
	Negation	82	22	8	52
	Non-monotone	30	17	7	6
	Temporal	32	11	11	10
	Universal	18	5	7	6
	Upward monotone	34	19	15	0
Knowledge	**Common sense**	150	36	56	58
	World knowledge	134	39	63	32

Table 4: GLUE diagnostics coarse- and fine-grained phenomena of language phenomena.