SweDiagnostics: A Diagnostics Natural Language Inference Dataset for Swedish

Felix Morger
University of Gothenburg
felix.morger@gu.se

Abstract

This paper presents SweDiagnostics, a natural language inference dataset for Swedish based on the GLUE Diagnostic dataset. It is the largest, manually corrected NLI dataset in Swedish to date and can be used to evaluate models on NLI in Swedish as well as estimate English-Swedish language transfer capabilities. We present the dataset, the methodology used for translation, compare existing implementations and discuss limitations of the dataset, in particular those related to translationese.

1. Introduction

Natural language inference (NLI) is the task of determining the logical relationship between two sentences. More specifically, whether a hypothesis entails, is neutral to or contradicts a given premise. For example, the hypothesis "John walks down the street" entails the premise "John is moving", but contradicts the premise "John is sitting" and is neutral to the premise "John is listening to music". NLI datasets have been created and studied extensively in natural language processing (NLP), based on the assumption that inferential reasoning is needed for all kinds of NLP tasks, such as question-answering, reading comprehension and sentiment analysis.

For English, several NLI datasets have come out, most notably the Multi-Genre Natural Language Inference (MultiNLI) (Williams et al., 2018) and Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015), which have 570K and 433K sentence pairs respectively. For cross-lingual evaluation, the Cross-lingual Natural Language Inference (XNLI) dataset extends these datasets with a separate 5K test split and 2.5K validation split of sentence pairs using the same data collection procedure, but is also translated into 15 different languages (Conneau et al., 2018). While these larger datasets were produced to provide enough training data to train large deep learning models, they are usually contrasted with smaller NLI datasets aimed to evaluate some specific phenomenon of interest (Poliak, 2020)[p. 94], usually called *test suites* or *diagnostics*. These consist of handpicked examples meant to target some phenomenon of interest. For example, Winogender (Rudinger et al., 2018) targets gender pronoun resolution and FraCaS (Cooper et al., 1996) covers a range of semantic phenomena from generalized quantifiers to temporal reference.

In this paper, we present a Swedish post-edited translation of the GLUE Diagnostic dataset. To date, this is the largest, manually corrected NLI dataset in Swedish (1106 samples), surpassing the only other Swedish NLI dataset SweFraCaS, which has 305 samples. This dataset allows for evaluating large language models (LLMs) on NLI for Swedish and English-Swedish language transfer similar to those done with the XNLI dataset. Such evaluation can be done by using existing English-Swedish machine translated NLI datasets as training data, such as those in Superlim and Overlim (Kurtz, 2022), an entirely English-Swedish machine translated version of SuperGLUE (Wang et al., 2019b).

The rest of the paper is structured as follows: We give an overview of the creation of the dataset (Section 2), compare existing implementations (Section 3), discuss potential limitations (such as with translationese and post-editese) and the development of future NLI datasets for Swedish (Section 4).

The dataset is available on HuggingFace as part of the Superlim project[1] and independently on the Språkbanken website.[2]

2. Dataset description

The GLUE Diagnostic dataset, which SweDiagnostics is based on, was released with the original SuperGLUE (Wang et al., 2019). It was handcrafted by linguistic experts with the aim to create a dataset for diagnosing a system's ability to solve a wide variety of language phenomena. The idea is to construct a hypothesis/premise sentence pair, where the entailment relationship depends on one or more targeted phenomena of interest. Table 1 illustrates this with two examples from the dataset. In the top sentence pair, the only difference be-

[1] https://huggingface.co/datasets/sbx/superlim-2

[2] https://spraakbanken.gu.se/resurser/swediagnostics

BUCC 2024: The 17th Workshop on Building and Using Comparable Corpora, pages 118–124
20 May, 2024.

tween the sentences is the added negation "did not" in the premise, which causes a contradiction. In the sentence pair below, the only difference is the added word "quietly" after "whispering" which is redundant since "whispering" (in most cases) implies talking quietly. Since the sentences express the same thing (i.e. talking quietly), the premise entails the hypothesis.

By way of this setup, NLI is used as a proxy to analyze specific language phenomena (negation and redundancy in the examples given). If the system can correctly predict the entailment relationship between the hypothesis/premise sentence pair, the conclusion is that the system encodes the targeted phenomenon.

The GLUE Diagnostic dataset has 33 different fine-grained language phenomena organized into four different coarse grained categories: lexical semantics, predicate-argument structure, logic and knowledge. Although the entailment relationship usually hinges on one particular fine-grained category, a sentence pair can be annotated with more than one category if the phenomenon is present in the text. Table 3 in the Appendix gives an exhaustive list of these categories as well as how many times they have been annotated in the dataset. For a detailed description of these categories, we refer to the latest documentation on the SuperGLUE website.[3]

2.1. Translation methodology

To create the equivalent SweDiagnostics, the sentence pairs were first machine translated using the Google Translate API. They were then post-edited by a native speaker of Swedish with a Master's degree in linguistics (the author of this paper). Besides adapting the translations to sound fluent and coherent, the translator also strove to uphold the two following criteria.

1. The entailment relationship remains the same after translation.
2. The annotated language phenomena remain the same after translation.

Although these criteria could not be fulfilled for every category due to morphological differences, such as in expressing double negation, in general this was not a problem. This is because (a) Swedish and English are closely related languages and, thus, share many of the morphological and syntactical features which are used to construct the contrasting sentence pairs and (b) the majority of the targeted linguistic phenomena of GLUE Diagnostic dataset are high-level natural language understanding features, which are not dependent on the particularities of English grammar.

The choice of post-editing over translating from scratch was done for efficiency reasons (cf. Plitt and Masselot (2010); Daems et al. (2017)). During translation the translator had the option of adding notes to document ambiguous or difficult parts of translation. Only 6.7% included notes, indicating a generally light post-editing effort.

3. Implementations

At the time of writing, SweDiagnostics has been evaluated in two separate projects. Firstly, as SweDiagnostics is a part of Superlim (Berdicevskis et al., 2023) it has been evaluated on multiple language models, both monolingual Swedish models and multilingual models. Secondly, a more fine-grained analysis has been done by Morger (2023), comparing English-Swedish language transfer capabilities of Swedish monolingual and multilingual models. In both of these projects, an English-Swedish machine translated version of MultiNLI was used for training.

In the discussion below as well as in Table 2 and Figure 1, model names are shortened for space reasons with the following abbreviations: `mt` for the "megatron" model (Shoeybi et al., 2019), `sw` for "Swedish", `l` for "large", `c` for "cased" and `b` for "base".

Table 2 shows the results on SweDiagnostics of Berdicevskis et al. (2023). Non-neural, supervised machine learning models are clearly outperformed by LLMs. The highest performing one is the multilingual model `xlm-roberta-large` (Conneau et al., 2019), outperforming the largest monolingual Swedish model `KBLab/mt-bert-l-sw-c-165k` (Malmsten et al., 2020). These results suggest that the amount of Swedish training data does not translate into increased performance. `KBLab/mt-bert-l-sw-c-165k`, for example, was trained on 70GB of only Swedish training data while `xlm-roberta-large` on 2.5TB of which only 12GB is in Swedish. The discrepancy in performance could also be explained by the difference in trainable parameters and language modeling objective. The fact that the sentences are originally English sentences could make it easier for the multilingual `xlm-roberta-large` model (see discussion in Section 4).

The results by Morger (2023), as reported in Figure 4, further compare the original GLUE Diagnostic dataset to SweDiagnostics. They concluded that a complete English-Swedish language transfer can be achieved using the English-

[3] https://super.gluebenchmark.com/diagnostics

	Swedish	English	
P	Katten satt på mattan.	The cat sat on the mat.	contradiction (**negation**)
H	Katten satt **inte** på mattan.	The cat **did not** sit on the mat.	
P	Tom och Adam viskade i teatern.	Tom and Adam were whispering in the theater.	entailment (**redundancy**)
H	Tom och Adam viskade **tyst** i teatern.	Tom and Adam were whispering **quietly** in the theater.	

Table 1: Two premise (**P**) and hypothesis (**H**) sentence pair examples from SweDiagnostics. The outermost column indicates the entailment relationship between the sentences. The annotated linguistic phenomenon which determines the relationship is in parentheses and marked **bold** in the text.

Swedish machine translated dataset of MultiNLI (cf. `bert-b-c` on GLUE Diagnostic dataset (blue bar) and `KB/bert-b-sw-c (mt-sv)` on SweDiagnostics (orange bar)). However, training on the original English data and only relying on multilingual pretraining (`bert-b-ml-c`) did not reach the same level of performance. Comparing this to `xlm-rorberta-large` in Table 2, this gap could possibly be filled by pre-training on more Swedish data or having larger architectures, but this remains speculative until a complete comparison has also been made to the `xlm-roberta-large` fine-tuned on English-Swedish machine translated data.

Overall, the fact that no model achieves higher than 0.44 Krippendorff's α (see `KB/bert-b-sw-c (mt-sv)` in Figure 1) shows that this task is still difficult for Swedish LLMs. Only a score above 0.67 is considered moderate agreement between the predicted and golden labels (Marzi et al., 2024). However, LLMs have made great headway towards solving this task when compared to non-neural, supervised machine learning models, which have scores close to 0 (i.e. no agreement) (see Table 2).

Model	**Krippendorf's α**
xlm-roberta-large	**0.415**
KBLab/mt-bert-l-sw-c-165k	0.393
KBLab/mt-bert-b-sw-c-600k	0.363
KB/bert-b-sw-c	0.349
AI-Nordics/bert-l-sw-c	0.347
KBLab/bert-b-sw-c-new	0.338
xlm-roberta-base	0.318
NbAiLab/nb-bert-base	0.314
Decision tree	0.037
SVM	0.026
Random forest	0.010
Random	0.004
MajLab/Avg	-0.404

Table 2: Evaluation results on SweDiagnostics as reported in Berdicevskis et al. (2023). They are reported in Krippendorf's α (Krippendorff, 2011), the metric of choice for Superlim. These are the results on eight different pretrained language models (upper part of the table) and five non-neural machine learning models (lower part of the table).

4. Concluding remarks

This paper has presented SweDiagnostics, an NLI dataset for Swedish, which is a post-edited, manually corrected version of the GLUE Diagnostic dataset.

As we see it, this resource provides three main contributions. Firstly, given the scarcity of NLI datasets for Swedish, this resource is an important addition in order to get *any* insights into the performance on NLI in Swedish, in particular monolingual Swedish language models. This is especially important given the release of multiple new monolingual Swedish language models in recent years, such as `KB-BERT` (Malmsten et al., 2020) and `GPT-SW3` (Ekgren et al., 2023). However, as the original authors of GLUE Diagnostic dataset are careful to point out, GLUE Diagnostic dataset is a *test suite* and, thus, one should be careful not to generalize over all language usage as it does not attempt to represent a natural language distribution. Secondly, SweDiagnostics's parallelity to GLUE Diagnostic dataset enables the comparison of English-Swedish cross lingual representations, which complements other multilingual resources, most notably XNLI (Conneau et al., 2018), which does not include Swedish. Thirdly, given the annotation of language phenomena in the dataset (see Section 2), further comparison can be made on the performance between different linguistic categories.

Creating a new resource by machine translating and post-editing an existing resource has both advantages and disadvantages. One of the most obvious advantages is that it is a cheap and efficient way to create a new resource, while another advantage is the resulting parallel corpora, which enables a close comparison between the languages. A disadvantage is that the samples

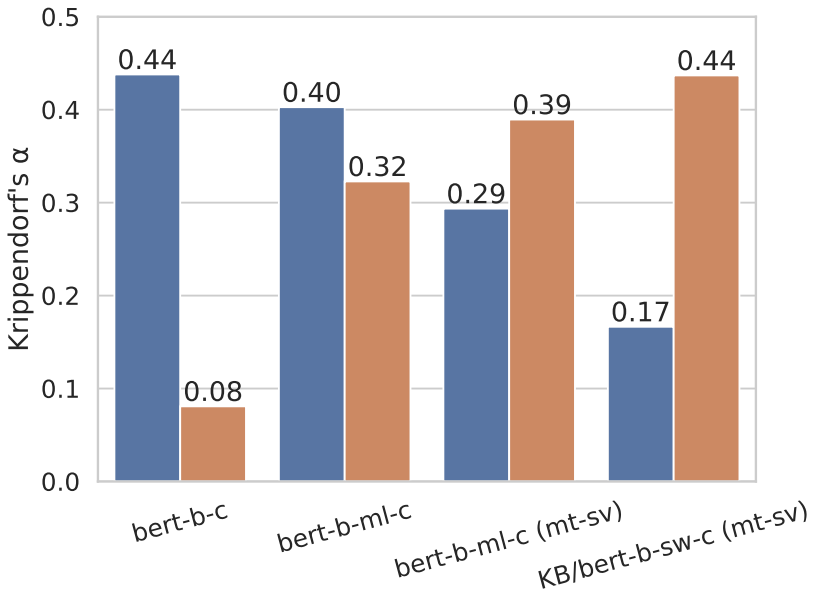

Figure 1: Model performance (Krippendorf's α) on the GLUE Diagnostic dataset (blue bars) and SweDiagnostics (orange bars) by (Morger, 2023). "mt-sv" refers to the model having been trained on the English-Swedish machine translated version of MultiNLI while the other ones are trained on the original English MultiNLI. Results can vary when compared with the original paper, which instead used the R_3 (Gorodkin, 2004) a three-class generalization of Matthews correlation coefficient (MCC).

are not taken from naturally occurring instances of the target language and will potentially not be a fair representation of the language overall. This is shown by Gellerstam (1986), which observe different statistical properties in translated language (translationese). This has also been shown to be further exacerbated by post-editing (Toral, 2019) (post-editese), however Daems et al. (2017) have shown that post-editing does not necessarily lead to lower quality translation. The results discussed in Section 3 do suggest that the performance on the GLUE Diagnostic dataset is highly transferable to SweDiagnostics, however, to what extent this is because of post-editese is unknown and could only be determined by future work systematically comparing post-editese to only human translations in the context of NLI.

This dataset together with SweFraCaS represents a first step towards evaluating NLI in Swedish. To get a fairer representation of Swedish and understand the effects of translationese, we encourage future work in creating new resources of NLI sourced from Swedish corpora. Comparing these to SweDiagnostics would not only give more insights into the NLI capabilities of Swedish monolingual and multilingual language models, but also insights into English-Swedish language transfer and language transfer between linguistically close languages more broadly.

5. Ethical considerations

As with any translated resource from a high-resource language to a lower-resource language, there is a risk of cultural biases being unfairly transferred to the target language. More broadly, using translated resources for evaluation could also amplify an anglocentric bias in what counts as the gold standard, which could divert funding from the creation of much needed unique language resources sourced directly from Swedish. For this reason, we encourage SweDiagnostics to be carefully compared with original Swedish resources and we also call for the creation of original NLI resources sourced exclusively from Swedish corpora.

6. Acknowledgments

This work was supported by Nationella språkbanken, which is jointly funded by the Swedish Research Council (2018–2024, grant no. 2017-00626) and 10 partner institutions.

7. Bibliographical References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Joke Daems, Sonia Vandepitte, Robert J Hartsuiker, and Lieve Macken. 2017. Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators. *Meta*, 62(2):245–270.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Alice Heiman, Judit Casademont, and Magnus Sahlgren. 2023. Gpt-sw3: An autoregressive language model for the nordic languages.

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.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. `https://www.asc.upenn.edu/sites/default/files/2021-03/ComputingKrippendorff'sAlpha-Reliability.pdf`.

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

Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-alpha calculator–krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545.

Felix Morger. 2023. Are there any limits to English-Swedish language transfer? a fine-grained analysis using natural language inference. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 30–41, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Mirko Plitt and François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague Bull. Math. Linguistics*, 93:7–16.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019b. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

8. Language Resource References

Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen, and Nina Tahmasebi. 2023. Superlim: A swedish language understanding evaluation benchmark. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page To appear, Sentosa, Singapore. Association for Computational Linguistics.

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 (EMNLP)*. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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.

Robin Kurtz. 2022. The KBLab blog: Evaluating Swedish language models.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana. 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*, volume 32. Curran Associates, Inc.

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. Association for Computational Linguistics.

Appendix

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