UNIVERSITY OF COPENHAGEN
DEPARTMENT OF NORDIC STUDIES AND LINGUISTICS
DEPARTMENT OF COMPUTER SCIENCE

Development and Evaluation of Pre-trained Language Models for Historical Danish and Norwegian Literary Texts

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen and Daniel Hershcovich

Abstract

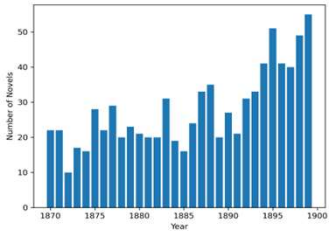
We develop and evaluate the first pre-trained language models specifically tailored for historical Danish and Norwegian texts. Three models are trained on a corpus of 19th century Danish and Norwegian literature: two directly on the corpus with no prior pre-training, and one with continued pre-training. To evaluate the models, we utilize an existing sentiment classification dataset, and additionally introduce a new annotated word sense disambiguation dataset focusing on the concept of "fate". Our assessment reveals that the model employing continued pre-training outperforms the others in two downstream NLP tasks on historical texts. Specifically, we observe substantial improvement in sentiment classification and word sense disambiguation compared to models trained on contemporary texts. These results highlight the effectiveness of continued pre-training for enhancing performance across various NLP tasks in historical text analysis.

Introduction

The wealth of digitized historical texts enriches research in Natural Language Processing (NLP) and Digital Humanities. Embedded with archaic terminology, context-sensitive variations, and cultural idiosyncrasies, these deviate vastly from contemporary texts. For instance, the evolution of the Danish term 'skæbne' illustrates how language changes over time, impacting perceptions of concepts like fate and responsibility. To tackle the challenges posed by historical texts, Pre-trained Language Models (PLMs) such as BERT have been trained on historical corpora across multiple languages, although resources for languages like Danish and Norwegian remain limited. In response, this study trains three PLMs on the Measuring Modernity (MeMo) corpus, consisting of Danish and Norwegian novels from 1870 to 1900, demonstrating improved adaptability in sentiment analysis and word sense disambiguation tasks compared to models trained on contemporary data. The aim is to deepen understanding of historical linguistic shifts in Danish and Norwegian, thereby empowering Digital Humanities researchers with enhanced tools and resources, which are made publicly available on HuggingFace.

Datasets

Main Corpus:
We rely on the MEMO corpus, comprising 839 Danish and Norwegian novels spanning the last 30 years of the 19th century and including more than 50 million words in total. The corpus is a rich and diverse collection of texts that will provide valuable insights into the registered sentiments and emotions of the period under investigation. The distribution of novels over years is shows in Figure 1.



Figur1: Distribution of Novels in the MeMo Corpus Over Time

Downstream Task Datasets:

Sentiment Analysis:
The first task we consider is sentiment analysis of sentences from historical Danish and Norwegian novels. A particularly suited dataset for our evaluation purpose is the sentiment classification dataset The dataset consists of 2,748 sentences from the MeMo corpus, annotated manually with three sentiment classes, negative, neutral and positive.

Sentiment Analysis			
Class	Samples	Percentage	
Negative	1,139	41%	
Neutral	788	29%	
Positive	821	30%	
	Training	**Validation**	**Testing**
	86%	10%	4%

Word Sense Disambiguation			
Class	Samples	Percentage	
Pre-modern	109	17%	
Modern	87	13%	
Figure of speech	275	42%	
Ambiguous	179	28%	
	Training	**Validation**	**Testing**
	70%	15%	15%

Table 1: Distribution of classes and training/validation/testing sets in both datasets, sentiment analysis and word sense disambiguation

Word Sense Disambiguation:
The second task we approach is word sense disambiguation in historical texts. In order to address it, we introduce a novel dataset established and annotated by one of the authors (a Danish-speaking literary scholar) investigating how the concept of fate "skæbne" is transformed in the latter part of the 19th century from its pre-modern sense, which is religiously and metaphysically inflected, to a modern meaning where the concept incorporates a secular and material understanding of the world. The dataset consists of 650 segments from the MeMo corpus, annotated manually with four classes, pre-modern, modern, figure of speech and ambiguous.

Table 2:F1-Score results of fine-tuning the selected models for sentiment analysis (SA) and word sense disambiguation (WSD) classification tasks on validation and test sets.

Task	SA		WSD	
Model	**Valid.**	**Test**	**Valid.**	**Test**
MeMo-BERT-1	0.52	0.56	0.41	0.43
MeMo-BERT-2	0.58	0.59	0.44	0.35
MeMo-BERT-3	**0.78**	**0.77**	**0.55**	**0.61**
DanskBERT	0.75	0.76	0.52	0.46
Danish BERT BotXO	0.74	0.74	0.19	0.30
ScandiBERT	0.73	0.73	0.40	0.40
DanBERT	0.65	0.63	0.39	0.41

MeMo-BERTs Models

MeMo-BERT-1:
The model is an instance initialized with the Transformer architecture, adhering to the BERT configuration. It comprises 12 layers, a hidden dimension of 768, 12 attention heads, and a vocabulary size of 30,000. This model is designed for tasks involving historical Danish and Norwegian text analysis, benefiting from its architecture's capacity to capture contextual information and semantic nuances.

MeMo-BERT-2:
The model diverges from MeMo-BERT-1 by adopting the XLM-RoBERTa architecture, offering enhanced depth and capacity. With 24 layers, a hidden dimension of 1024, 16 attention heads, and a subword vocabulary size of 50,000, this model exhibits heightened performance potential. Its architecture enables more nuanced understanding of complex linguistic features present in historical Danish and Norwegian texts.

MeMo-BERT-3:
*I*n contrast to the first two models, utilizes the pre-trained Transformer PLM DanskBERT, which is specialized for the contemporary Danish Gigaword Corpus. It inherits its architecture from XLM-RoBERTa, boasting 24 layers, a hidden dimension of 1024, and 16 attention heads. Moreover, it features a significantly larger subword vocabulary size of 250,000, indicative of its focus on handling a broader range of contemporary Danish language nuances.

For all models, we use the masked language modelling objective in pre-training, where 15\% of the input tokens are masked, and the model is trained to predict the original tokens. The models are all encoder-only Transformers with case sensitivity. The corpus is randomly split into 80\% for training and 20\% for validation. We set the batch size for training to 16 and validation to 32, the number of gradient accumulation steps to 8, the learning rate to 1e-4, the number of training epochs to 3, the maximum number of training steps to 12500, and the number of warm-up steps to 1250. We select the best checkpoint based on validation loss. These parameters have a significant impact on the convergence and performance of the trained model. The training process is performed in a distributed manner, utilizing two A100 GPUs, and takes 44, 36, and 32 hours for training the three models respectively.

Experiments

To evaluate the developed historical models and the baselines trained on contemporary Danish (\S\ref{sec:baselines}), we use two downstream tasks, sentiment analysis and word sense disambiguation. These tasks were selected on the basis of their relevance for historical text processing. Both are represented by datasets annotated over text from the same MeMo corpus that we use for pre-training the PLMs. We select the comparison models based on their popularity and accuracy in similar tasks. The models were tested on diverse NLP benchmark datasets \cite{nielsen-2023-scandeval}. We utilize them to assess the performance of our developed historical models.

Results and Discussion

Table 2 shows that F1-Score performance of various models in Sentiment Analysis (SA) and Word Sense Disambiguation (WSD) tasks, highlighting MeMo-BERT-3's superior performance over MeMo-BERT-1, MeMo-BERT-2, and other models in both SA and WSD tasks.

MeMo-BERT-3's exceptional performance in both SA and WSD tasks, emphasizing its superiority over MeMo-BERT-1, MeMo-BERT-2, and other models. It also mentions DanskBERT's competitive performance, attributed to its pre-training on a mix of contemporary and historical Danish text.

the effectiveness of MeMo-BERT-3 in capturing nuanced linguistic features through pre-training on historical text, corroborated by its superior performance over other models in both SA and WSD tasks. It also highlights the beneficial impact of incorporating historical language data on model comprehension and classification accuracy for text from historical periods.

Conclusion

In this research, we introduce the first PLMs for historical Danish and Norwegian, trained on the MeMo corpus of 839 novels from the late 19th century, outperforming models trained on contemporary texts.

Our future plans include expanding training data with historical documents from various periods, assessing model generalizability across diverse historical corpora, and developing annotated datasets for tasks like named entity recognition and event extraction to enable more nuanced literary analysis.