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

Department of Computer Science
Department of Nordic Studies and Linguistics

KØBENHAVNS UNIVERSITET

Agenda

- Introduction
- Datasets
- Models' Development
- Models' Evaluation
- Results and Analysis
- Conclusion

Introduction

- The increasing of digitized historical texts enhances research in Natural Language Processing (NLP) and Digital Humanities.

- Pre-trained Language Models (PLMs) like BERT are employed to address challenges posed by historical texts, although resources for languages like Danish and Norwegian are limited.

Introduction

- This study trains three PLMs on the Measuring Modernity (MeMo) corpus, comprising Danish and Norwegian novels from 1870 to 1900, showcasing improved adaptability in sentiment analysis and word sense disambiguation tasks compared to models trained on contemporary data.

- The objective is to deepen understanding of historical linguistic shifts in Danish and Norwegian, thereby empowering Digital Humanities researchers with enhanced tools and resources.

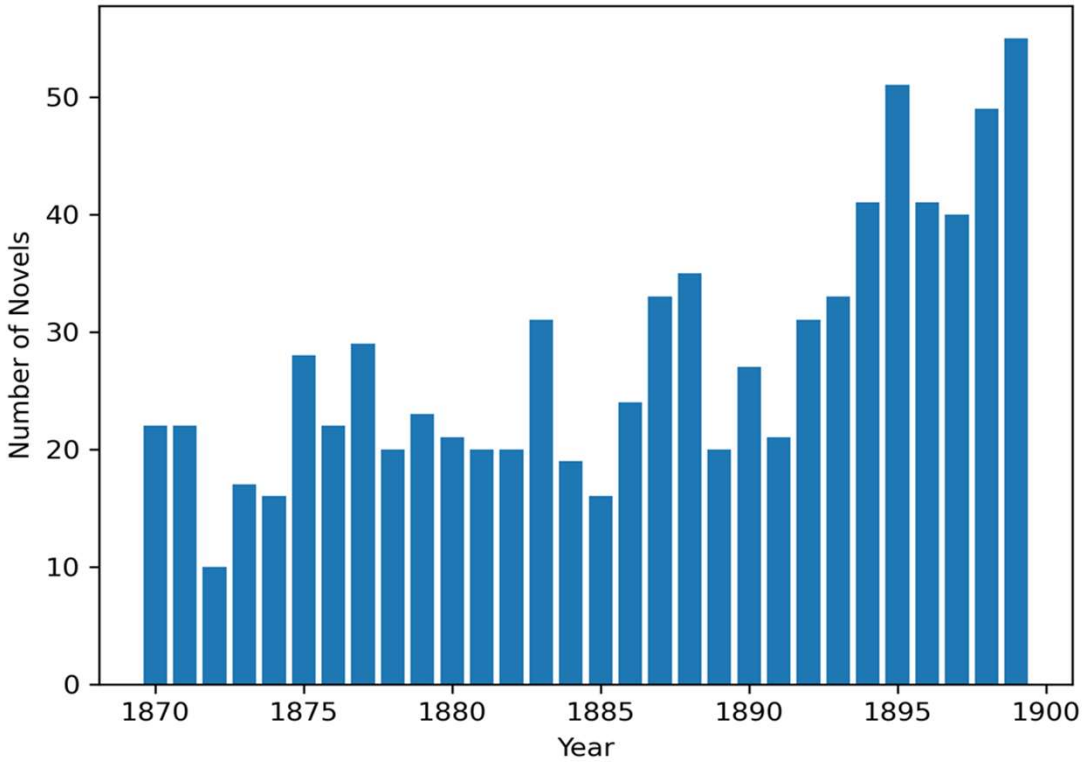
Datasets

- **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.

Datasets

- **Main Corpus:**



Datasets

- **Downstream Task Datasets:**

1. 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.
- These sentences are annotated manually with three sentiment classes: negative, neutral, and positive.

Datasets

- **Downstream Task Datasets:**

2. 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).
- The dataset investigates 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.
- These segments are annotated manually with four classes: pre-modern, modern, figure of speech, and ambiguous.

Datasets

- **Downstream Task Datasets:**

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%

Models Development

- We introduce MeMo-BERTs, three PLMs designed to support historical Danish and Norwegian text.

- We employ two different approaches to train the models:
  1. One approach involves training the models directly on the MeMo Corpus.
  2. The other approach utilizes continued pre-training based on historical and contemporary-language PLM.

Models Development

- **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.
  - It benefits from its architecture's capacity to capture contextual information and semantic nuances.

Models Development

- **MeMo-BERT-2:**
  - The model diverges from MeMo-BERT-1 by adopting the XLM-RoBERTa architecture.
  - It offers enhanced depth and capacity compared to MeMo-BERT-1.
  - 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 a more nuanced understanding of complex linguistic features present in historical Danish and Norwegian texts.

Models Development

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

Models Development

- **Pre-training Setup:**
  - For all models, we use the masked language modelling objective in pre-training.
  - In this objective, 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.

Models Development

- **Pre-trining Setup:**
  - We set the batch size for training to 16 and for validation to 32.
  - The number of gradient accumulation steps is set to 8.
  - The learning rate is set to 1e-4.
  - The number of training epochs is set to 3.
  - The maximum number of training steps is set to 12500, and the number of warm-up steps is set to 1250.

Models Development

- **Pre-training Setup:**
  - We select the best checkpoint based on F1-score.
  - These parameters significantly impact the convergence and performance of the trained model.
  - The training process is performed in a distributed manner, utilizing two A100 GPUs.
  - The training duration for the three models is 44, 36, and 32 hours respectively.

Models Evaluation

- **Experiments:**
  - To evaluate the developed historical models and the baselines trained on contemporary Danish, we use two downstream tasks: sentiment analysis and word sense disambiguation.
  - These tasks were selected based on their relevance for historical text processing.
  - Both tasks are represented by datasets annotated over text from the same MeMo corpus used for pre-training the PLMs.

Models Evaluation

- **Experiments:**
  - We select the comparison models based on their popularity and accuracy in similar tasks.
  - The models were tested on diverse NLP benchmark datasets.
  - We utilize them to assess the performance of our developed historical models.

Results and Analysis

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

Results and Analysis

- The F1-Score performance of various models in Sentiment Analysis (SA) and Word Sense Disambiguation (WSD) tasks highlights 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 emphasizes its superiority over MeMo-BERT-1, MeMo-BERT-2, and other models.

- Additionally, DanskBERT's competitive performance is mentioned, attributed to its pre-training on a mix of contemporary and historical Danish text.

Results and Analysis

- The effectiveness of MeMo-BERT-3 in capturing nuanced linguistic features through pre-training on historical text is highlighted.

- Moreover, the beneficial impact of incorporating historical language data on model comprehension and classification accuracy for text from historical periods is emphasized.

Conclusion

- In this research, we introduce the first Pre-trained Language Models (PLMs) for historical Danish and Norwegian, trained on the MeMo corpus.
- One of the developed models outperforms models trained on contemporary texts.
- Our future plans include expanding training data with historical documents from various periods.

Conclusion

- We aim to assess model generalizability across diverse historical corpora.
- Additionally, we plan to develop annotated datasets for tasks like named entity recognition and event extraction.
- These efforts aim to enable more nuanced literary analysis using the developed PLMs.

Thank you!

Contact:

Ali Al-Laith
alal@di.ku.dk