SLaLaM 2023
1st Slovenian workshop on Large Language Models Techniques and Applications

Bernardin, Slovenia, December 21, 2023

Workshop proceedings

Edited by Jaya Caporusso and Nada Lavrač
Jožef Stefan Institute, Ljubljana, Slovenia

Workshop Agenda and Proceedings Table of Contents:

FIRST SESSION: Pre-LLM news analysis research

1. **Senja Pollak (Jožef Stefan Institute): EMBEDDIA news analysis**

We present automated news analysis methods that were developed in the scope of the European project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). More specifically, we focus on cross-lingual technologies and applications for keyword extraction, diachronic analysis, comment filtering and news generation. We discuss the applicability of the methods for the media monitoring industry.

2. **Matthew Purver (Jožef Stefan Institute and Queen Mary University of London): Cross-lingual comment filtering**

We present work from the EMBEDDIA and RobaCOFI projects, analysing comments under news articles to detect those that should be moderated, approached as text classification with cross-lingual transfer learning. Focus on less-resourced languages, including Croatian and Slovenian; training classifiers on intermediate data in other well-resourced languages including English and German. This worked really well, giving reasonable zero- and few-shot performance, and better accuracy with full training data than a monolingual equivalent; but required a custom-built en/hr/si language model. However, we then showed that this requirement can be relaxed by a good choice of multilingual pre-trained LLM, and using unsupervised domain adaptation; our models and code are available. Some work on humans in the loop, including active learning for few-shot improvements, and on classifier decision explanation/visualisation.

3. **Marko Pranjić (Jožef Stefan Institute and Jožef Stefan International Postgraduate School): News clustering pilot for EMMA project**

The presented work deals with the challenging task of detecting groups of related news articles in a multilingual setting with varying numbers and sizes of article groups. With the prevalence of individual news stories covered by multiple media houses, each offering a distinct perspective, our goal is to identify related groups of news articles that provide a comprehensive overview of a given topic.

To tackle this problem, we use SentenceTransformers (SBERT) to generate article embeddings. The multilingual capabilities and strong results on semantic tasks of SBERT makes it a good choice. To detect groups, we use Louvain community detection on the binarised cosine similarity matrix generated from the SBERT representations.

We use metrics such as Silhouette score, Rand index, Mutual Information and Pairwise F1 score, which provide a comprehensive insight into the performance of our solution. Although we achieve promising results and improve the baseline approaches, our approach is limited by the maximum sequence length of the model, the non-overlapping nature of the resulting clusters, as well as the inherent instability of the community detection algorithm resulting in small differences in results across multiple runs. These results contribute to the ongoing discourse on effective methods for clustering news articles and provide a solution that combines the strengths of pre-trained multilingual models and graph-based clustering.

4. **Boshko Koloski (Jožef Stefan Institute and Jožef Stefan International Postgraduate School): Neuro-symbolic zero-shot cross-lingual keyword detection**

In our study, we evaluate the effectiveness of supervised and unsupervised zero-shot cross-lingual methods in a multilingual environment. We discovered that embedding-based methods surpass traditional unsupervised approaches, such as statistical and graph-based techniques. Notably, in cross-lingual contexts, zero-shot transfer from one language to another demonstrates superior performance compared to unsupervised methods. However, we observed a performance decline when more languages were added to the training set. Zero-shot methods show greater efficiency for cross-lingual tasks, whereas models specifically adapted to individual tasks perform better in single-language scenarios. Integrating traditional methods like TF-IDF with large language models presents a potent approach, particularly effective in keyword detection tasks.

5. **Nikola Ivačič (Dropchop, Jožef Stefan Institute, Jožef Stefan International Postgraduate School): Zero-Shot Detection of News Framing Dynamics in European Migration Crises: A Multilingual Transformer Approach**

This study presents an approach to understanding the evolution of news framing in the context of two significant European migration crises: the Syrian-Middle East conflict and the Ukraine war, as portrayed in the Slovenian media. Utilising a zero-shot learning framework, we explore the divergence in news framing between these two periods.

Initially, our work involved fine-tuning the pre-trained multilingual BERT and XLM-roBERTa models using the multilingual corpus, annotated with five news-framing labels (Economy, Labour Market, Welfare, Security, Culture) across seven European languages. This corpus, derived from the EU REMINDER project, comprised both automatically and manually annotated samples, totalling over 60,000 entries.

Our primary objective was to construct and analyse a target language (Slovene) corpus specific to the migration periods of interest and make a zero-shot prediction. The Slovene corpus was sourced from the Clipping Industry archive, focusing on migration-related keywords and covering data sets from August 2015 to April 2016 (Syria-Middle East crisis) and February 2022 to March 2023 (Ukraine crisis).

At the heart of our approach, we focused on training two multilingual models based on pre-trained neural network Transformer architectures. We employed both Binary Relevance and Label Power-set techniques for transforming the problem at hand. These models were then evaluated against a small corpus that had been manually annotated. We tested 8 different model combinations in a 10-fold cross-validation process.

When applying the model to Slovene, we observe that the economy and the security frame are much more frequent in the Syrian than in the Ukrainian period.

SECOND SESSION: Current LLM research at JSI

6. **Nikola Ljubešić (Jožef Stefan Institute and University of Ljubljana) and Taja Kuzman (Jožef Stefan Institute and Jožef Stefan International Postgraduate School): Specialization of LLMs and their benchmarking across the South-Slavic language continuum (and beyond)**

In our research, we address the evolving landscape of Large Language Models (LLMs) with a focus on specialization and benchmarking across the South-Slavic language continuum and beyond. Transitioning from traditional language model training where the models are developed for a less-resourced language by pre-training from scratch, as part of the MaCoCu project, we experimented with a more pragmatic approach. We started from the massively multilingual XLM-RoBERTa language model and additionally pretrained it on HBS

(Croatian, Serbian, Bosnian and Montenegrin) and Slovenian data, leading to the creation of XLM-R-BERTić and XLM-R-SloBERTić models. We showcase that this approach provides similar or better performance on named-entity recognition and sentiment regression tasks than pretraining from scratch, while being quicker and less computationally expensive. However, when the models are evaluated on a commonsense reasoning task, their performance deteriorates, highlighting challenges related to a drift from shared multilingual spaces. Similarly, as part of the ParlaMint project, we produced a model specialized for parliamentary discourse, the XLM-R-parla model. We evaluate it on the task of sentiment regression using the multilingual ParlaSent dataset, manually annotated with sentiment, and show minor, but consistent improvements obtained with additional pre-training, and large benefits of multilingual training.

Expanding our scope to cross-lingual genre classification, we present the X-GENRE classifier. The classifier is based on multilingual XLM-RoBERTa model which was fine-tuned on English-Slovenian manually-annotated genre datasets. We highlight the model's high zero-shot cross-lingual performance on related languages and minimal impact of scripts on performance. Interestingly, when the classifier is applied on a completely unrelated language that was not included in the XLM-RoBERTa pretraining data, namely, Maltese, the model's performance significantly deteriorates, nevertheless, it recognized some of the labels with near-perfect accuracy.

Finally, our research touches on datasets relevant for the news analysis research, including massive web MaCoCu corpora, annotated with genre labels, allowing extraction of news content, and the Trendi monitor corpus of Slovenian news which is automatically annotated with a high-performing topic classifier. We conclude with insights into the performance of instruction-tuned large language models on a commonsense reasoning benchmark, COPA. While the models reach very high performance on South Slavic languages, we show that the dialects are more challenging for the LLMs. Lastly, we introduce yugoGPT – a grassroots initiative, developing a South Slavic LLM, pretrained on our HBS data. The model demonstrates promising performance in the COPA-SR benchmark compared to existing models.

7. **Boshko Koloski (Jozef Stefan Institute and Jožef Stefan International Postgraduate School): Topic identification with AHAM**

We introduce AHAM, a novel topic modeling validation framework for literature-based discovery. AHAM integrates Large Language Models (LLM), particularly LLAMA, with Generative Pseudo Labeling (GPL) guided by expert-driven prompt engineering. This method effectively identifies and refines topic clusters, employing GPL for domain adaptation. A key innovation is using LLAMA, guided by prompt engineers, to generate descriptive and relevant topic names. The framework focuses on reducing outliers and enhancing topic distinctiveness. Our evaluation demonstrates AHAM's effectiveness in creating coherent, domain-specific topics, surpassing traditional approaches.

8. **Hanh Thi Hong Tran (University La Rochelle, Jožef Stefan Institute and Jožef Stefan International Postgraduate School): LLM prompting for named entities and term extraction**

We present our research on the applicability of open and closed-sourced large language models (LLMs) on the ATE task in comparison with two benchmarks where we consider ATE as sequence-labeling (iobATE classifier) and seq2seq ranking tasks (templATE classifier), respectively. The aim is to bridge the gap between text generation and sequence labeling tasks by guiding the models to produce predictions with three designed formats. We present the prompt design, including task description, few-shot demonstration, and input sentence. We compared different approaches to LLM prompting (IOB format, original gold standard format, and generative format) and compared the results on the ACTER dataset in three languages. Our empirical inquiry unveils that LLMs' prompting performs close to fully supervised sequence-labeling baselines, and offers a valuable trade-off by eliminating the need for extensive data annotation efforts.

THIRD SESSION: Current LLM research at FRI

9. **Aleš Žagar (University of Ljubljana): Summarization, document representation, document visualization**

We presented how summarization can be viewed along the following dimensions: based on input type (single or multi-document), based on output type (extractive or abstractive), and based on purpose (general and query-based are the most relevant for this project). Initially, summarization research predominantly utilized unsupervised techniques, such as prioritizing sentences based on keyword frequency or part-of-speech significance. However, machine-learning approaches have evolved to consider various sentence features, including position and keyword density. Recently, the seq2seq models that are build on Transformer architectures, produce the most impressive summaries. Some short text summarization datasets exist (STA; AutoSentiNews, and automatically translated CNN/DailyMail dataset). The problem with these datasets is that they do not contain human-written abstracts, therefore the first paragraph is used as an approximation of the summary. For long text summarization, we have a high-quality corpus of Slovene academic texts (KAS). We also created four summarization models that are based on various approaches, and a metamodel to suggest which model should be the most appropriate based on the specific text of interest. To encode the input document, we used a Doc2Vec model. Additionally, we introduced a visualization tool that graphically represents sentence importance, connectivity, and semantic similarity, enhancing understanding and analysis of text summarization processes.

10. **Tadej Škvorc (University of Ljubljana): Text clustering, conference scheduling, merging text presentations**

We present an approach for conference scheduling based on embeddings obtained from text and network-based metadata of scientific papers. We cluster research papers using embedding vectors generated from text content (word2vec, fasttext, Elmo/BERT vectors and terminology extraction) and network-based metadata obtained from citation and review co-bidding graphs, which are then concatenated into the final representation vectors. We evaluate which features are useful using feature-selection approaches on a separate classification task. We use the relevant features to cluster papers with similar meanings and arrange papers so that papers with similar contents do not appear in different tracks at the same time. We evaluate our approach on multiple datasets mimicking real-world conferences and show that our approach is able to minimise scheduling conflicts between papers with similar topics.

11. **Martin Božič (University of Ljubljana): Approaches for language correction**

We divide grammar correction tasks into separate categories and for each category synthetically generate test and train datasets. For each category we then fine-tune several models and sequentially connect best performing models into the pipeline. These categories include: word order correction, comma placement correction, incorrect word spelling annotation, incorrect word spelling correction and syntax error correction. Finally we incorporate all models into an intuitive web application, where each grammar correction category is labelled with a different colour, so the user knows what sorts of mistakes the application corrected.

12. **Domen Vreš (University of Ljubljana): SloLLaMa: the Slovenian large language model for text generation**

As a part of the project PoVeJMo we are training Slovene large language models for text generation. Our plan is to train 2 general models - the smaller one with 1B parameters (OPT model will be used as a starting point) and larger one with around 10B parameters (either LLaMa 2 or Mistral will be used as a starting point). We will train the models using Nvidia's NeMo toolkit. As there is not enough Slovene text to train the models from scratch, the pre-trained weights of English models will be preserved. Both models will be equipped with a new tokenizer that is trained on Slovene, Croatian, Serbian, Bosnian and English texts. We will initialise the new embedding matrices by transferring the embeddings of already trained English models to new languages. In the second part of the project, we will fine-tune the models for instruction following and specific industrial use cases.

13. **Marko Robnik-Šikonja (University of Ljubljana): Cross-lingual transfer and model interpretation**

The techniques of cross-lingual transfer nowadays mainly rely on multilingual Large Language Models (LLMs). For text analytics and generation, the most useful are transformer LLMs such as Croatian/Slovene/English CroSloEnglual BERT, massively multilingual mBERT, and even larger generative LLMs like ChatGPT, Orca-2, T5-XXL, and LLaMA-2. The alternative to the model-based transfer is transfer using machine translation of datasets. The transfer is based on three phases: pretraining (usually already done), fine-tuning in well-resourced language, and using the model in the target language (zero-shot transfer), optionally using few-shot transfer. The main problem with massively multilingual LLMs are vocabularies, where the tokens in a dictionary depend on all languages, and more languages mean that the individual tokens are less semantic. The dictionary is constructed statistically from input corpora.

The explainability with LLMs is problematic, but maybe we can learn something from humans, who often cannot explain their decisions. Here LLMs correspond to System1 reasoning in Khaneman's book Thinking fast, thinking slow. The goal is to produce explanations akin to System2, which is logical and step-wise. One shall be aware that explanations are problem-dependent. We can divide the explanation options based on the existence of background knowledge (well-structured, unstructured, inexistent).

FOURTH SESSION: Current JSI and FRI research in representation learning for media applications (Moderators: Nada Lavrač, Marko Robnik-Šikonja) - out of proceedings scope