SHORT TERM SCIENTIFIC MISSION (STSM)
SCIENTIFIC REPORT

This report is submitted for approval by the STSM applicant to the STSM coordinator

Action number: CA18209 - European network for Web-centred linguistic data science
STSM title: Machine learning for detecting and interpreting language phenomena in survey data
STSM start and end date: 30/08/2021 to 09/09/2021
Grantee name: Kostadin Mishev

PURPOSE OF THE STSM:

(max.200 words)
The STSM was carried out within WG4, UC 4.2.2 and its main purpose is to exchange experience and know-how on using Machine Learning (ML) methods, such as contextual word embeddings from transformer models and distributed word representations for detecting and interpreting language phenomena in texts from survey data, TED talks, and social media in English, Lithuanian, Bulgarian, German, Macedonian and Portuguese. The outcome of the STSM is providing various methodologies based on Natural Language Processing and eXplainable AI for analysis and interpretation of the results of a series of experiments over these three categories of datasets, and the extraction of semantic information, related to discourse markers (DM), attitude expressions, opinion detection, and sentiment from them. Verification of the transfer-learning methods of DM detection in Discovery dataset and their application to the annotated sets in English, Bulgarian and Lithuanian. Evaluation of using the language-agnostic and cross-lingual models and building a tool for semi-automatic labeling.
The main contribution of the STSM to the scientific objectives of NexusLinguarum Action is to provide a framework for linguistic data science that addresses some prototypical cases of interpretation of language phenomena with social significance using the latest advents in semantic representation, transfer-learning and eXplainable AI.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

(max.500 words)

The following tasks were carried out during the STSM:

1. Review of the available datasets and methods applied over the datasets for extracting the language phenomena. The analysis included a discussion about the dataset format, sentence representation, labels analysis, and statistical analysis.
2. Preparation of different state-of-the-art methodologies used in general tasks and adapt them for detecting and interpreting language phenomena in survey data, including recognition of discourse markers, sentiment extraction, attitude expression, and opinion detection. Validation of the chosen methodologies with the host.
3. Performing data preprocessing on the datasets, including tokenization, lemmatization, stemming, stop-words removal. Due to the multilinguistic nature of the proposal, we will consider using language-agnostic and cross-lingual preprocessing methods.

4. Preparation of scripts for classification NLP models that we decided to use for our tasks. Unsupervised domain data model training and supervised fine-tuning depending on the task. Due to the cross-lingual nature of the study, we started the analysis with distributed word representation incorporating subword information using the FastText word embeddings and LASER sentence embeddings. On the one hand, FastText provides cross-lingual embeddings, and on the other hand, LASER provides language-agnostic sentence embeddings, so we found them suitable for this task. The text-classification model includes convolutional units combined with attention layers. Next, we developed the study to include the latest NLP transformers, including XLM-RoBERTa-large, representative of cross-lingual language models, and La-BSE, as language-agnostic sentence encoder. First, we had to pretrain them on the domain data, and next, we fine-tuned them by adding an additional layer that performs text classification.
5. The setup of the previous models was applied for the discourse marker detection text classification task. English, Bulgarian and Lithuanian text corpora were already annotated, so we performed separate evaluations per dataset. Each of the datasets was split into training and testing datasets using the ratio 75%:25% accordingly. The models were trained/fine-tuned on the training set, and the evaluation was performed on the test set. Afterward, the cross-lingual and language-agnostic nature of the task was observed. Each of the cross-lingual models, FastText and XLM-RoBERTa, was trained/fine-tuned on the English dataset and evaluated on the datasets from the other languages to examine the ability to transfer the knowledge obtained in the English language and applied in the different languages whose corpus were already annotated: Bulgarian and Lithuanian. Finally, the language-agnostic nature of discourse marker detection was explored by training LASER and La-BSE as sentence encoders in the English dataset and applied to the Bulgarian and Lithuanian datasets.
6. A multi-class classification method was developed and evaluated on the Discovery dataset. The model used in the evaluation was XLM-RoBERTa-Large.
7. After evaluating the language-agnostic methods, the La-BSE model trained on the English dataset was leveraged to aid in annotating the unannotated corpus, including German, Portuguese, and Macedonian datasets.

The SHAP, a method from eXplainable AI (XAI), was used to explain the cross-lingual and language-agnostic model decisions when identifying the discourse markers in the sentence.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

The Corpora are TED talks, a parallel corpus containing data from 6 languages, using the publicly available TED Talk transcripts. It is an ongoing expansion of TED-EHL parallel corpus published in LINDAT/CLARIN-LT repository http://hdl.handle.net/20.500.11821/34. The multilingual corpus contains alignments of Lithuanian, Bulgarian, Portuguese, Macedonian, and German languages with English as pivot language with a size of 1.3 million sentences. Secondly, we constitute a vocabulary of multiword expression that can play the role of discourse markers in text based on theoretical insights by Schiffrin (1987) and classification provided by Fraser(2009). The next step was the manual annotation of the 2428 English-Bulgarian-Lithuanian aligned sentences containing the multiword expressions (MWE) as discourse markers or content expressions (1 or 0). Probably we will publish the parallel corpus of the 6 languages on CLARIN.

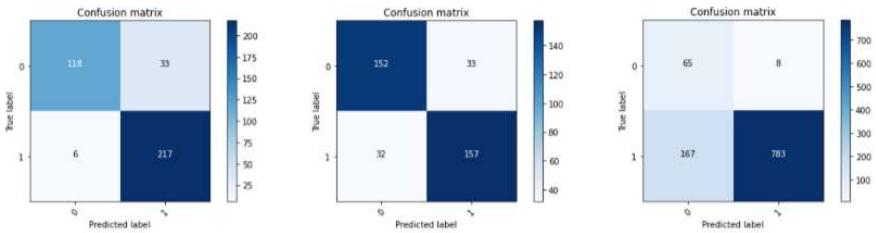
Due to the binary nature of the task, Discourse Marker (DM) detection in a sentence, as the primary evaluation metrics, we choose Matthews Correlation Coefficient (MCC). MCC is widely used in assessing binary classification performance with a range between -1 (completely wrong classifier) and 1 (completely accurate classifier). Additionally, it provides a balanced measure since it takes into consideration true and false positives and negatives.

Table 1. presents the evaluation results achieved using the cross-lingual methods, FastText and XLM-RoBERTa-large, applied on DM detection in a sentence, when all text-corpus were considered separately per language. Each of the datasets was split into train and test using the ratio 75%:25% accordingly. Figure 1. visualizes the results achieved by XLM-RoBERTa-large using a confusion matrix.

The results prove the power of the contextual XLM-RoBERTa-large's embeddings since they outperform the distributed word embeddings provided by FastText in the evaluated languages (EN, BG, LT). Furthermore, it means that transformer architecture is more appropriate for DM detection tasks. This result can be considered a reference in future research by the workgroup.

Model	Accuracy	Precision	Recall	Specificity	F1-Score	MCC
FastText (EN)	0.4558	0.6515	0.1928	0.8467	0.2976	0.0507
FastText (BG)	0.5764	0.6457	0.6457	0.4733	0.6457	0.1191
FastText (LT)	0.9321	0.9369	0.9942	0.0548	0.9647	0.1285
XLM-RoBERTa (EN)	0.918	0.890	0.786	0.913	0.903	0.808
XLM-RoBERTa (BG)	0.826	0.826	0.830	0.822	0.829	0.652
XLM-RoBERTa (LT)	0.8289	0.9899	0.8242	0.8904	0.8995	0.4393

Table 1. Results achieved on datasets in different languages using cross-lingual methods



(a) English (b) Bulgarian (c) Lithuanian

Figure 1. Confusion Matrix of XLM-RoBERTa-large results

In Table 2, the results using language-agnostic methods are presented. In this case, we fine-tuned Facebook's LASER and Google's La-BSE, both representatives of language-agnostic sentence representation models, with the English dataset and evaluated on the Bulgarian and Lithuanian datasets. La-BSE showed better results compared to the LASER encoder. It is worth noting that the set of discourse markers is different for each language, but in this research, the language-agnostic models extracted only the inherent set of DMs for the pair of languages, and the methods from XAI helped us to discover this conclusion.

Model	Accuracy	Precision	Recall	Specificity	F1-Score	MCC
LASER (BG)	0.628	0.652	0.567	0.654	0.631	0.382
LASER (LT)	0.742	0.876	0.832	0.231	0.843	0.082
La-BSE (BG)	0.727	0.740	0.709	0.745	0.724	0.455
La-BSE (LT)	0.834	0.941	0.876	0.288	0.907	0.123

Table 2. Language-Agnostic Methods Results

Additionally, we harnessed the language-agnostic model La-BSE, which outperformed Facebook's LASER in the previous step, to develop a semi-automatic annotation tool. This tool will support the linguists to easily annotate the unannotated text-corpus in English, Bulgarian, Lithuanian, German, Macedonian, and Portuguese with the appropriate label for DM presence. The number of automatically annotated sentences that we supported by developing the tool is given in Table 3.

Language	Number of sentences
English	33.028
Bulgarian	17.399
Lithuanian	2.982
German	15.851
Macedonian	2.845
Portuguese	4.398

Table 3. The number of unannotated sentences labeled automatically with the developed language-agnostic method for DM detection.

Our work presents the ground work to go towards extracting discourse markers and representing in LLOD the discourse relations, the semantic and pragmatic information they provide to the communication.

The methods and codes developed during the STSM are publicly available at the following link: https://github.com/f-data/DM-Detection

FUTURE COLLABORATIONS (if applicable)

The methodology and results developed in the research during this STSM are robust and present a novelty in methods for cross-lingual DM detection. Thus, they establish a base for the publication of a scientific paper in a conference or journal. We extended the network with another linguist from Majka Teresa University in Skopje, who will help us enlarge the Macedonian corpora and build a novel Albanian corpus. With my host institution Mozaika, we plan to continue our collaboration in this field in the future. During the stay, we discussed that our methodology could be harnessed to build up a study that clusters natural languages by similarities in the discourse marker sets, thus finding correlations within the language families.