Research and Innovation Action (RIA) H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu

D2.6 Intermediate Release of Segmentation, Summarization and News Classification Capabilities

Work Package	2
Responsible Partner	Priberam
Author(s)	Diogo Pernes
Contributors	Afonso Mendes, Tugtekin Turan
Reviewer	Tugtekin Turan
Version	1.0
Contractual Date	31 December 2022
Delivery Date	22 December 2022
Dissemination Level	Public

Version History

Version	Date	Description
0.1	03/11/2022	Introduction and Initial Table of Contents (ToC)
0.2	15/12/2022	Internal Review Version
0.3	19/12/2022	Ready for the Final Review
1.0	22/12/2022	Publishable Version

Executive Summary

This intermediate deliverable describes the second release of software components developed within WP2 and integrated with the SELMA orchestration platform. These components mainly apply to Use Case 1 (UC1), the “Media Monitoring Platform”.

SELMA targets creating stream-based models that can leverage updated news information to improve the topic classification. Using SELMA’s monitoring platform, users can keep track of current trending news stories across multiple languages.

This document provides an overview of the release of the “Online Multilingual News Clustering”, “Topic Detection”, “News Summarization” and Story Segmentation capabilities to follow the final releases later in the project. We highlight the extended multilingual capabilities of the new releases.

Table of Contents

Table of Figures

1. Introduction

This intermediate deliverable describes the cumulative release of software components developed within WP2 and integrated with the SELMA orchestration platform. These components are mainly applicable to use case 1 (UC1), the "Media Monitoring Platform" except for the segmentation component which is relevant to both UC1 and UC2. We address components of the Natural Language Processing (NLP) document enrichment pipeline, namely, "Multilingual Topic Classification", "Online Multilingual New Clustering", "Document summarization" and segmentation. This report aims to describe those components from the point of view of the release of software components. This deliverable should be read in conjunction with D2.4, where a complete technical description is reported.

All components are deployed as Docker containers (www.docker.com) and will be made available at https://hub.docker.com/orgs/selmaproject, expose REST APIs and provide swagger documentation pages (https://swagger.io). These components are integrated with the SELMA orchestration platform and are already being used by UC1 (https://app.monitio.com).

2. Released Components

2.1 Online Multilingual News Clustering

Multilingual News Clustering is a core piece of the Media Monitoring use case as it allows users to focus on stories instead of being overwhelmed by a massive amount of scattered news articles. Our work for the SELMA platform was presented in D2.1 and later accepted at the Text2Story Workshop held at ECIR 2022. Unlike other enrichment tasks like Topic Classification or NER/NEL, news clustering is not easily scalable to handle a very big incoming stream of documents and thus can be a bottleneck in the processing pipeline. The reason for this is that the clustering of a document depends on the status of the clustering pool at a certain point in time, making it impossible to scale by adding additional workers. The component's performance is thus an essential aspect when integrating into the pipeline.

The currently deployed version can process 3.2 documents per second on average while maintaining a cluster pool of 25,000. Given that the platform is currently ingesting about

150,000 documents per day, this seems an acceptable performance; nevertheless, since the flow is not uniform during the day, we already see times when there is a perceivable delay between the ingestion and the clustering. Research efforts are being carried out in WP2 to prevent this problem. The current model handles 50 languages *(en, ar, bg, ca, cs, da, de, el, es, et, fa, fi, fr, fr-ca, gl, gu, he, hi, hr, hu, hy, id, it, ja, ka, ko, ku, lt, lv, mk, mn, mr, ms, my, nb, nl, pl, pt, pt-br, ro, ru, sk, sl, sq, sr, sv, th, tr, uk, ur, vi, zh-cn, zh-tw),* as described in D2.1. This model is *state-of-the-art* in the task of online multilingual news clustering as reported in D2.1 and in our publication at the Text2Story workshop "Simplifying News Clustering Through Projection from a Shared Multilingual Space" by João Santos, Afonso Mendes, and Sebastião Miranda.

The clustering engine is deployed as a docker container that exposes a REST API which exposes the following methods:

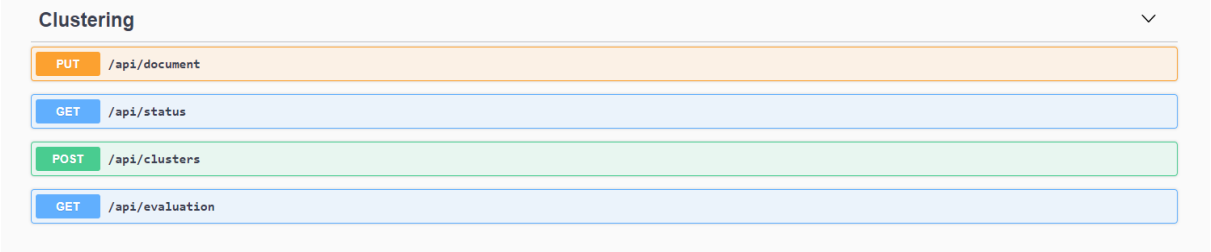

Figure 1 *NER Annotation Tool*

The first method /api/document receives a document and returns the clustering ID to associate the document with, the method also returns pairs of (doc id, cluster id) when previously clustered documents, due to a cluster merge, are reassigned after the current operation. The /api//status and /api/clusters are for state monitoring purposes, and the /api/evaluation evaluates the algorithm as a sanity check before deployment.

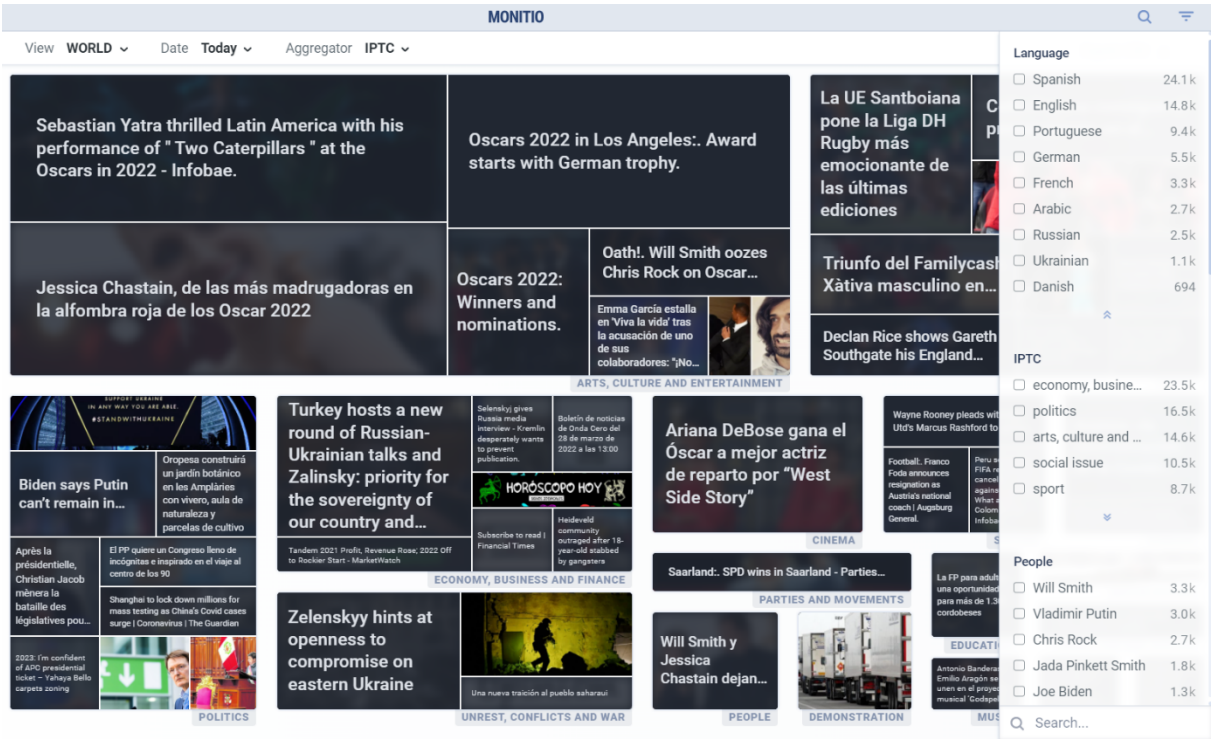

***Figure 2** "Storylines" Dashboard from Monitio Showing Multilingual Clustering*

During the second reporting period this component has been updated in order to improve its performance. Namely, the communication pipeline has been changed to use gRPC instead of REST and the embeddings creation has been decoupled from the module.

2.2 Topic Detection

In SELMA, we have two objectives regarding topic detection and news classification. The first is to enhance the multilingual capabilities of the IPTC Subject Codes classification defined by the International Press Telecommunications Council. The second is the ability of the system to classify documents against user-defined topics given a minimum amount of user feedback.

During the reporting period, as reported in D2.1, we have developed and released a new multilingual model for the IPTC topic classification task based on the AttentionXML (You et al. 2019) that improves F1 performance in the test sets for Portuguese, English, and Spanish by 5.8%, 3,6%, and 11,8%, respectively, against our previous SUMMA model. Even though the model was not yet formally evaluated in other languages, initial user feedback on the accuracy for other languages like Russian and Arabic is very good; we expect a formal user evaluation

to start soon in the scope of WP5. The model was trained to leverage the contextual embeddings of mBERT, which was trained in 102 languages.

The model was deployed as a docker container exposing the same REST API as the old SUMMA model for compatibility reasons and integrated into the SELMA orchestration platform.

While research efforts continue on the topic of few-shot classification and active learning, we have created and deployed the UX for the “Smart tags” scenario as described in D1.2 together with a baseline model for few-shot classification. These will allow us to collect user data for evaluation.

During the second reporting period we deployed a new version of the classification module trained with more data and released the new explainability features as described in D2.4. The previous version of the model showed strange behaviors when dealing with very small documents and that was corrected by introducing a second model on the pipeline to deal with those cases. This is now integrated on UC1 with a good evaluation by the platform users.

2.3 News Summarization

SELMA proposes to advance the state of the art in abstractive summarization by tackling the problem of factually inconsistent and irrelevant summaries and by extending current research to multi- and cross-lingual settings. While the latter problem is the subject of future work, the former was approached with a re-ranking approach as described in Pernes et al. (2022) and reported in D2.4.

The whole summarization pipeline was deployed as a docker container that includes both the abstractive summarizer (BART- or Pegasus-based) and the re-ranking model, which was fine-tuned from BERT. The available API exposes several parameters that allow fine-grained control over the summary generation process. The interface is shown in Figure 3. Currently, it can only perform English-to-English summarization. The release of a summarization component for cross-lingual summarization is within our plans as soon as our research on this problem allows it.

Figure 3 *Abstractive Summarization API*

2.4 Story Segmentation

In this release, our goal is to segment speakers accurately to identify the different speakers in an audio recording by assigning them to distinct labels. This can be useful for various applications, such as transcribing audio recordings, summarizing the content of a conversation, or analyzing the structure of a conversation. Therefore, identifying who spoke when can be regarded as a preliminary task to the story segmentation process. Our current approach to this problem is to use a hidden Markov model (HMM) to map the sequence of speech segments belonging to each speaker.

In our proposed approach, speaker embeddings are used as a compact representation of the acoustic characteristics of a speaker's voice. These embeddings are fixed-dimensional feature vectors extracted from a deep neural network trained for speaker recognition, where feature vectors capture information about the spectral and prosodic characteristics of the speaker's voice, and they can be used to discriminate between different speakers.

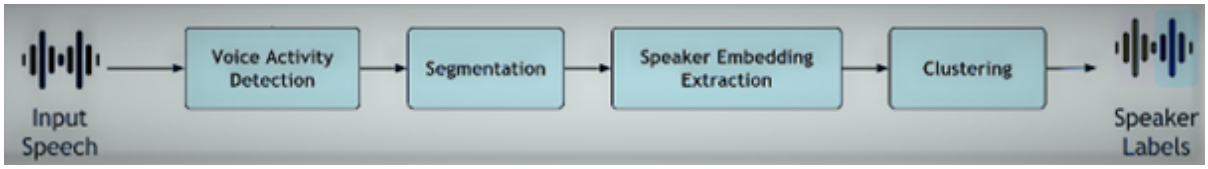

Figure 4 *An Overview of the Proposed Speaker Segmentation Pipeline*

To perform speaker segmentation using HMMs, we first need to extract embeddings from the audio segments of interest (e.g., short time frames of speech). Next, we can use a clustering algorithm to group them into clusters, with each cluster representing a different speaker. One way to do this clustering is with a Bayesian HMM, a type of HMM that uses Bayesian inference to estimate the model's parameters.

In a Bayesian HMM, each speaker is represented by a separate HMM, and the speaker embeddings are assumed to be observations generated by these HMMs. The parameters of the HMMs (e.g., the transition probabilities between states) are estimated using Bayesian inference, allowing uncertainty to be incorporated into the model. This uncertainty can be helpful when there is limited training data, as it makes the model more robust to variability in the data.

At the final stage, to cluster speaker embeddings using a Bayesian HMM, we would use an iterative algorithm (e.g., the expectation-maximization (EM) algorithm) to estimate the parameters of the HMMs and assign each embedding vector to the HMM that is most likely to have generated it. The resulting clusters can then label the corresponding audio segments of different speakers.

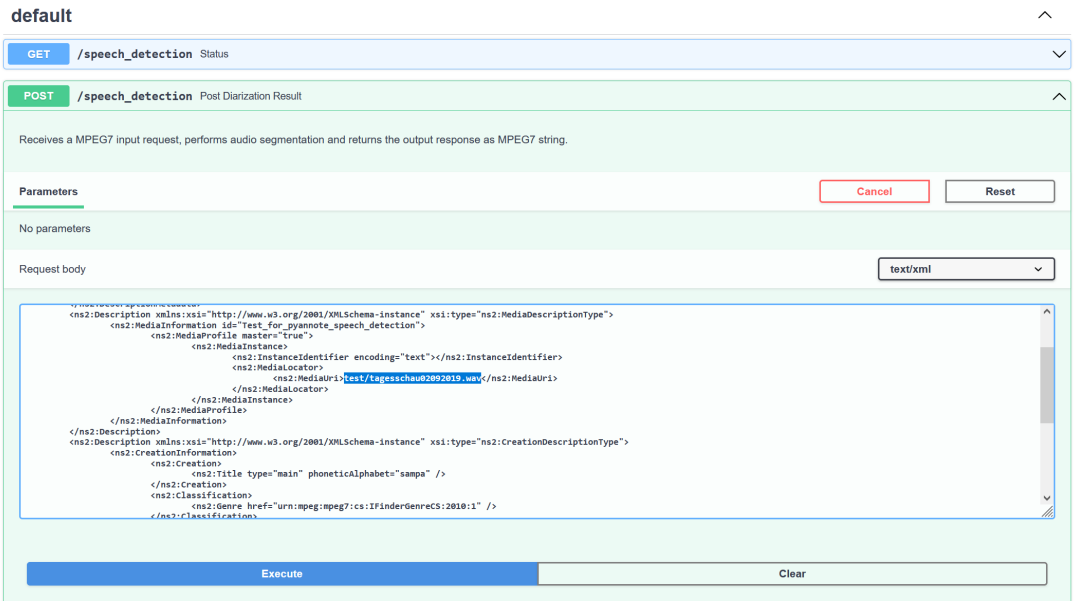

Figure 5 *Speaker Segmentation API*

This approach can be used independently of linguistic context, as it attempts to cluster the speaker rather than the content of the conversation. The speaker segmentation interface is shown in Figure 5, which is structured to handle the whole pipeline depicted in Figure 4. The real-time factor (RTF) is a standard metric for measuring the speed of the segmentation process. Our API provides an RTF of 4, in which for every 1 hour of recording, the proposed system has a processing time of approximately 15 minutes.

3. Future Plans

This software release covers the WP2 tasks of T2.3, T2.4, and T2.5. We successfully integrated a multilingual news classification module. We significantly extended the language support of the models with better production scalability and better performance. We made available the first speaker segmentation module which is language agnostic.

For the subsequent releases of software components, our focus will be:

1) Continue to improve clustering performance,
2) Release of the first components for multilingual and cross-lingual summarization,
3) Release of the first models for topic classification using user feedback.

Bibliography

Barzdins, G., Gosko, D., Cerans, K., Barzdins, O. F., Znotins, A., Barzdins, P. F., Gruzitis, N., Grasmanis, M., Barzdins, J., Lavrinovics, I., Mayer, S. K., Students, I., Celms, E., Sprogis, A., Nespore-Berzkalne, G., Paikens, P. (2020b). Pini Language and PiniTree Ontology Editor: Annotation and Verbalisation for Atomised Journalism. *In: ESWC 2020 Satellite Events. LNCS, Volume 12124, pp. 32-38.*

Peteris Paikens; Guntis Barzdins; Afonso Mendes; Daniel Ferreira; Samuel Broscheit; Mariana S. C. Almeida; Sebastiao Miranda; David Nogueira; Pedro Balage; Martins, Andre F. T. (2016a). SUMMA at TAC Knowledge Base Population Task 2016, *DOI: 10.5281/zenodo.827317*

Znotiņš, Artūrs & Barzdins, Guntis. (2020). LVBERT: Transformer-Based Model for Latvian Language Understanding. *Baltic HLT, IOS Press, pp. 111-115, DOI 10.3233/FAIA200610.*

Znotins A, Cirule E. (2018). NLP-PIPE: Latvian NLP Tool Pipeline. *In: Human Language Technologies - The Baltic Perspective. vol. 307. IOS Press; 2018. p. 183–189.*

Paikens P. (2016b). Deep Neural Learning Approaches for Latvian Morphological Tagging. *In: Baltic HLT; 2016. p. 160–166.*

Pernes, D., Mendes, A., & Martins, A. F. (2022). Improving abstractive summarization with energy-based re-ranking. *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2022)*.

J Santos, J., Mendes, A. & Miranda, S. (2022). Simplifying Multilingual News Clustering Through Projection From a Shared Space. *Proceedings of Text2Story - Fifth Workshop on Narrative Extraction From Texts* held in conjunction with the *44th European Conference on Information Retrieval (ECIR 2022)* Stavanger, Norway, April 10, 2022 (pp. 015-024)

You, R., Zhang, Z., Wang, Z., Dai, S., Mamitsuka, H., & Zhu, S. (2019). AttentionXML: Label Tree-based Attention-aware Deep Model for High-performance Extreme Multi-label Text Classification. *Advances in Neural Information Processing Systems*.