Research and Innovation Action (RIA) H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu/

D2.8 Final release of continuous massive stream learning tools

Work Package	2
Responsible Partner	Priberam
Author(s)	Diogo Pernes
Contributors	Afonso Mendes, Gonçalo Correia
Reviewer	Andreas Giefer
Version	V1.0
Contractual Date	31 March 2024
Delivery Date	28 March 2024
Dissemination Level	Public

Version History

Version	Date	Description
0.1	03/03/2024	Updated Table of Contents (ToC)
0.2	13/03/2024	Input other partners
0.3	17/03/2024	Internal Review
1.0	25/03/2024	Publishable version

Executive Summary

This deliverable cumulatively describes the final release of software components developed within WP2 and integrated with the SELMA orchestration platform. These components are mostly applicable to 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. SELMA's approach to named entities is to continuously learn new named entities from the reference stream and link them to a knowledge base (e.g., Wikipedia).

This document provides an overview of the final releases of the "Online Multilingual News Clustering", "Topic Detection", "News Summarization", "Story Segmentation", "Stream Learning", and "Entity Linking" capabilities of the project. We highlight the extended multilingual capabilities of the released components.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

This final deliverable cumulatively describes the release of software components for stream learning and entity linking developed within WP2 and the integration with the SELMA orchestration platform. These components are mostly applicable to UC1, the Media Monitoring Platform.

We address several components of the Natural Language Processing (NLP) document enrichment pipeline, such as named entity recognition (NER), multilingual named entity linking (NEL), online multilingual news clustering, topic detection, single and multi-document summarization, and story segmentation. This report's objective is to describe those components from the point of view of the deployment of software components, meaning its characteristics, requirements and specifications at the time of deployment. This deliverable should be read in conjunction with D2.7, where a full technical description is reported.

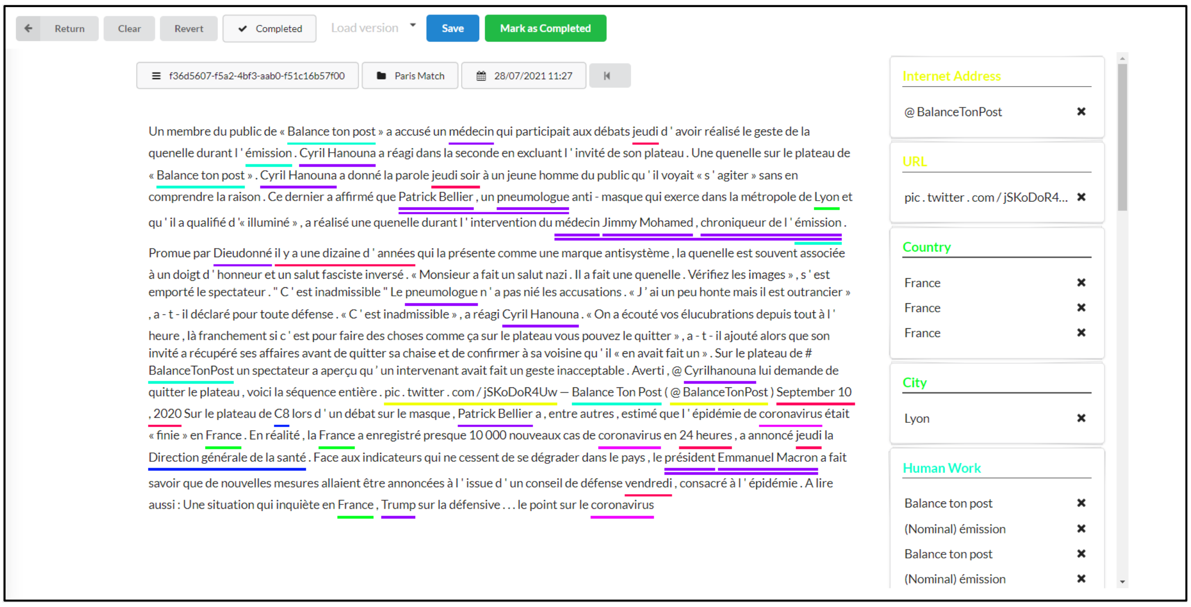
All components are deployed as Docker containers (www.docker.com) and were made available at https://hub.docker.com/orgs/selmaproject. They expose REST APIs and provide swagger documentation pages (https://swagger.io/). These components are integrated with the SELMA orchestration platform and are already being used by Use Case 1 (UC1) (https://app.monitio.com).

2. Released components

2.1 Annotation tool

An annotation tool for Named Entity Recognition was developed by Priberam and deployed at https://www.priberam.com/annotate. It is used by Priberam, IMCS, and DW for the development of a multilingual named entity dataset available in several languages. Priberam has already annotated news documents in Portuguese (3000 documents), French (3000 documents), Spanish (in progress), German (3000 documents), and English (6000 documents). IMCS annotated Latvian, Russian and Ukranian. DW annotated Dutch and Turkish.

The annotation tool was extended for the task of Entity Linking, thus allowing the annotator to link a mention to a specific Entity in the knowledge base, (e.g linking Obama to the Entity Q76 for Barack Obama on Wikidata).



f36d5607-f5a2-4bf3-aab0-f51c16b57f00 | Paris Match | 28/07/2021 11:27

Un membre du public de « Balance ton post » a accusé un médecin qui participait aux débats jeudi d ' avoir réalisé le geste de la quenelle durant l ' émission . Cyril Hanouna a réagi dans la seconde en excluant l ' invité de son plateau . Une quenelle sur le plateau de « Balance ton post » . Cyril Hanouna a donné la parole jeudi soir à un jeune homme du public qu ' il voyait « s ' agiter » sans en comprendre la raison . Ce dernier a affirmé que Patrick Bellier , un pneumologue anti - masque qui exerce dans la métropole de Lyon et qu ' il a qualifié d '« illuminé » , a réalisé une quenelle durant l ' intervention du médecin Jimmy Mohamed , chroniqueur de l ' émission .

Promue par Dieudonné il y a une dizaine d ' années qui la présente comme une marque antisystème , la quenelle est souvent associée à un doigt d ' honneur et un salut fasciste inversé . « Monsieur a fait un salut nazi . Il a fait une quenelle . Vérifiez les images » , s ' est emporté le spectateur . " C ' est inadmissible " Le pneumologue n ' a pas nié les accusations . « J ' ai un peu honte mais il est outrancier » , a - t - il déclaré pour toute défense . « C ' est inadmissible » , a réagi Cyril Hanouna . « On a écouté vos élucubrations depuis tout à l ' heure , là franchement si c ' est pour faire des choses comme ça sur le plateau vous pouvez le quitter » , a - t - il ajouté alors que son invité a récupéré ses affaires avant de quitter sa chaise et de confirmer à sa voisine qu ' il « en avait fait un » . Sur le plateau de # BalanceTonPost un spectateur a aperçu qu ' un intervenant avait fait un geste inacceptable . Averti , @ Cyrilhanouna lui demande de quitter le plateau , voici la séquence entière . pic . twitter . com / jSKoDoR4Uw — Balance Ton Post (@ BalanceTonPost) September 10 , 2020 Sur le plateau de C8 lors d ' un débat sur le masque , Patrick Bellier a , entre autres , estimé que l ' épidémie de coronavirus était « finie » en France . En réalité , la France a enregistré presque 10 000 nouveaux cas de coronavirus en 24 heures , a annoncé jeudi la Direction générale de la santé . Face aux indicateurs qui ne cessent de se dégrader dans le pays , le président Emmanuel Macron a fait savoir que de nouvelles mesures allaient être annoncées à l ' issue d ' un conseil de défense vendredi , consacré à l ' épidémie . A lire aussi : Une situation qui inquiète en France , Trump sur la défensive . . . le point sur le coronavirus

Internet Address
- @ BalanceTonPost ✖

URL
- pic . twitter . com / jSKoDoR4... ✖

Country
- France ✖
- France ✖
- France ✖

City
- Lyon ✖

Human Work
- Balance ton post ✖
- (Nominal) émission ✖
- Balance ton post ✖
- (Nominal) émission ✖

***Figure 1** NER Annotation tool*

2.2 Named Entity Recognition models using HNNNER

During the first reporting period, we deployed a set of trained models using the datasets developed in the scope of SELMA according to the annotation guidelines defined in D6.1. In particular, we deployed Spanish, English, Portuguese, German and French and we are evaluating whether the model trained on the already available Latvian dataset achieves the minimum requirements to be deployed on UC1.

As described in D2.1, we developed two different models, the stack-LSTM and the Biaffine model for Hierarchical Nested Named Entity Recognition. Both models presented similar F1 scores in our news test Datasets, where the stack-LSTM has shown better performance (processing speed) characteristics. Therefore, this model was selected for deployment in the UC1 scenario.

These models have been deployed as a docker container exposing a REST API which is used by the SELMA orchestration platform. The docker containers were deployed on machines with GPUs, so that the document processing throughput was compatible with UC1 requirements. The complete NLP pipeline used on UC1 is currently able to ingest and process 150 000 documents per day.

The NER REST API is composed by a single method that receives a JSON object, where the text is provided as shown in following example:

```
POST /predict/ HTTP/1.1
Host: pbacomp03.interno.priberam.pt:8800
Content-Type: text/plain
Content-Length: 202
{
  "text": "Susan Kersch-Kibler, Gründerin der Agentur Delivering Dreams, hat ihre
Leihmütter kurzerhand ins Ausland verfrachtet – nur um sie zum Geburtstermin
wieder in die Ukraine zu befördern. "
}
```

It produces the following output, where for each entity a JSON object is returned using the following format:

mention_id	Sequential id
mention	Mention text
total_offset	Character Offset from the beginning of input text
end_offset	Character End offset from the beginning of text
sentence_offset	Near context offset form the beginning of text
near_context	Text sentence where the mention was detected
ner_tag	NER type as defined in the Ontology
ner_type	‘NAM’ for named entity or ‘NOM’ for nominal entity
ner_modifiers	Modifiers applied to the NER type like collective, negated etc. (see ontology in D6.1)

```
[
    {
        "mention_id": 0,
        "mention": "Susan Kersch-Kibler",
        "length": 19,
        "total_offset": 0,
        "end_offset": 19,
        "sentence_offset": 0,
        "near_context": "Susan Kersch-Kibler, Gründerin der Agentur Delivering
Dreams, hat ihre Leihmütter kurzerhand ins Ausland verfrachtet – nur um sie zum
Geburtstermin wieder in die Ukraine zu befördern. ",
        "sentence_id": 0,
        "ner_tag": "people",
        "ner_type": "NAM",
        "ner_modifiers": []
    },
    {
        "mention_id": 1,
        "mention": "Delivering Dreams",
        "length": 17,
        "total_offset": 43,
        "end_offset": 60,
        "sentence_offset": 0,
        "near_context": "Susan Kersch-Kibler, Gründerin der Agentur Delivering
Dreams, hat ihre Leihmütter kurzerhand ins Ausland verfrachtet – nur um sie zum
Geburtstermin wieder in die Ukraine zu befördern. ",
        "sentence_id": 0,
        "ner_tag": "organization->commercial_company",
        "ner_type": "NAM",
        "ner_modifiers": []
    },
    {
        "mention_id": 2,
        "mention": "Agentur Delivering Dreams",
        "length": 25,
        "total_offset": 35,
        "end_offset": 60,
        "sentence_offset": 0,
        "near_context": "Susan Kersch-Kibler, Gründerin der Agentur Delivering
Dreams, hat ihre Leihmütter kurzerhand ins Ausland verfrachtet – nur um sie zum
Geburtstermin wieder in die Ukraine zu befördern. ",
        "sentence_id": 0,
        "ner_tag": "organization->commercial_company",
        "ner_type": "NAM",
        "ner_modifiers": []
    },
    …
```

Multilingual Named Entity Recognition

During the last reporting period we deployed a new fully multilingual version of the NER component. The component supports 100 languages, the same as the base model xlm-roberta-base. The component was trained on English, German, Spanish, French, Latvian, Russian, Ukrainian, and Portuguese. It was evaluated on unseen languages during the training in Dutch, Ukrainian and Turkish and showed a very good zero-shot performance. Qualitative evaluation leads us to believe that the component zero-shots with very reasonable results to many of the other languages. It is now integrated in UC1, and the users' perception is very good across all languages represented in the platform. Details about the evaluation for this component can be found in D2.4. Besides the language support, the model allows us to better scale the processing pipeline since we only use the resources (CPU, GPU, RAM) for one model instead of one per language.

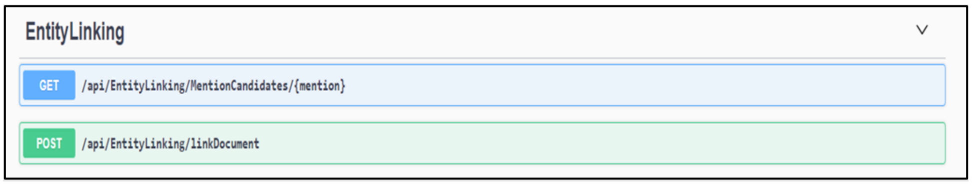
The deployment details are the same as the previous NER component.

2.3 Multilingual Entity Linking with Wikidata/Wikipedia as the Knowledge Base

The first version of the Entity Linking models described in D2.1 were deployed for UC1 in the Monitio platform. The model deployed during that period supported the same languages as the NER models described in the previous section (Spanish, Portuguese, German and French) covering a total of 14 322 317 different entities. Since Wikidata and Wikipedia are constantly being updated, and new entities inserted, we have implemented an automatic procedure to collect and incrementally update the knowledge base data. This procedure incrementally trains and updates the representations for the entities as described in D1.2.

The model was deployed as docker container exposing a REST API. A single instance can process 2 documents per second on average when deployed over GPU. To cope with the current MONITIO stream, we have currently deployed two instances of the service.

The REST API exposes two methods: one to obtain the possible candidates for a given mention and another to perform the actual Entity Linking (EL) for a given document.


EntityLinking

GET /api/EntityLinking/MentionCandidates/{mention}

POST /api/EntityLinking/linkDocument

Figure 2 *Swagger method definition for the EL component*

For each linked entity, the service provides the Wikidata identifier for the entity and the Wikipedia title in the original language and English. We include additional metadata regarding the entities like binary gender, date of birth, country, occupation, etc... as defined in the SELMA requirements for the diversity use case (see D1.1 - 3.1.1.2 Diversity).

Currently, the service returns for each entity (additional information is available in the swagger page):

baseForm	English Wikipedia title if available; if not, the same as currlangForm
currLangForm	Wikipedia title in the original language of the document
id	Wikidata unique identifier
type	one of people, location, gpe (geopolitical entity like countries), organization and event

Sample Json output from the EL service:

```
{
    "entities": [
        {
            "entity": {
                "baseForm": "Aung San Suu Kyi",
                "currlangForm": "Aung San Suu Kyi",
                "id": "Q36740",
                "lowest_confidence": 21.091115864011837,
                "type": "people"
            },
            "mentions": [
                {
                    "confidence": 1,
                    "endPosition": {
                        "chunk": 0,
                        "offset": 37
                    },
                    "ner_type": "people",
                    "sourceDocument": null,
                    "startPosition": {
                        "chunk": 0,
                        "offset": 21
                    },
                    "text": "Aung San Suu Kyi"
                },
                {
                    "confidence": 1,
                    "endPosition": {
                        "chunk": 0,
                        "offset": 296
                    },
                    "ner_type": "people",
                    "sourceDocument": null,
                    "startPosition": {
                        "chunk": 0,
                        "offset": 284
                    },
                    "text": "Aung Suu Kyi"
                },
….
```

Extended Multilingual Entity Linking

During the second reporting period, we deployed a new entity-linking component with support for 39 languages. The following table lists the languages supported by the EL model. The Technical description for this module can be found on D2.4.

Language Code	Language
am	Amharic
ar	Arabic
bg	Bulgarian
bn	Bengali
bs	Bosnian
ca	Catalan
cs	Czech
de	German
el	Greek
en	English
es	Spanish
fa	Farsi
fi	Finnish
fr	French
ha	Hausa
he	Hebrew
hi	Hindi
hr	Croatian
hu	Hungarian
id	Indonesian
it	Italian
ja	Japanese
lv	Latvian
mk	Macedonian
nl	Dutch
no	Norwegian
pl	Polish
ps	Pashto
pt	Portuguese
ro	Romenian
ru	Russian
sq	Albanian
sr	Serbian
sv	Swedish
sw	Swahili
tr	Turkish

uk	Ukrainian
ur	Urdu
zh	Chinese

***Table 1** EL Language support*

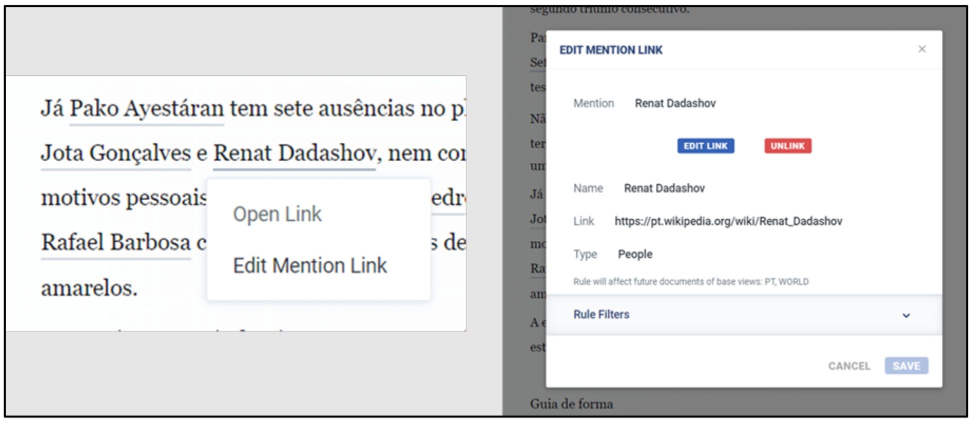
During the last reporting period, we deployed an EL model with the same multilingual support of 39 languages, but with an increase in number of entities to 20,000,000. This model has an increased performance on several test sets compared with the previous API. The test set results of the complete pipeline (NER, candidate selection, and entity disambiguation) are shown in Table 2, with a comparison between the previous API and the currently deployed one.

Test set	Previous API	Current API
Aida-B	0.7753	**0.8596**
TAC 2016 en-news	0.8868	**0.9213**
TAC 2016 en-discussion forums	0.8903	**0.8908**
TAC 2016 es-news	0.9215	**0.9243**
TAC 2016 es-discussion forums	0.884	**0.8885**
TAC 2016 zh-news	0.7746	0.7746
TAC 2016 zh-discussion forums	0.8257	**0.8636**
VoxEL en	0.9274	**0.9689**
VoxEL es	0.8969	**0.9226**
VoxEL de	0.8385	**0.8958**
VoxEL it	0.9119	**0.9585**
VoxEL fr	**0.9528**	0.9424

***Table 2** EL results*

NER/EL user feedback in the MONITIO platform

According to the requirements defined for UC1, the MONITIO platform was extended to gather NER / NEL corrections and additions.



Já Pako Ayestáran tem sete ausências no p
Jota Gonçalves e Renat Dadashov, nem co
motivos pessoais edr
Rafael Barbosa c s de
amarelos.

Mention Renat Dadashov

Name Renat Dadashov

Link https://pt.wikipedia.org/wiki/Renat_Dadashov

Type People

Rule will affect future documents of base views: PT, WORLD

Figure 3 *Feedback collection user interface for entity linking and NER*

The functionality is already being used by test users and the data is being collected.

2.4 Rule-Based Stream Learning for NEL (PiniTree Ontology Editor)

The SELMA partner IMCS, the University of Latvia, has been involved in the NEL topic for several years (Barzdins, 2020; Paikens, 2016a), jointly with the Latvian national news agency LETA and PiniTree.com startup. This has resulted in the development of the commercial PiniTree.com ontology editor with integrated rule-based Stream learning of Named Entity Linking aliases as part of the entity database, against which the Named Entities are being Linked. PiniTree editor is one of the tools being integrated into the SELMA Platform. In addition, the LETA use case (described in detail in deliverable D1.2) is available for wider exploitation along with other SELMA components. Within the initial release of the SELMA Platform, PiniTree is integrated into Use Case 0 as the backend content management system accessed via the “Publish” button.

Technically, PiniTree is a universal web server with integrated database and user management. PiniTree is distributed as a single precompiled binary file downloadable from https://pinitree.com, therefore no specific installation is required – on Linux, the PiniTree server is started directly by executing the PiniTree binary file from the command-line:

```
bash % ./pinitree.linux-amd64 -p 7777 -a
```

The rest of the PiniTree web server configuration takes place through the graphical web interface. The first time when trying to connect to the newly started PiniTree web server, it will display the following prompt to create the first (admin) user. Additional users with various privileges can be added later by logging in as admin.

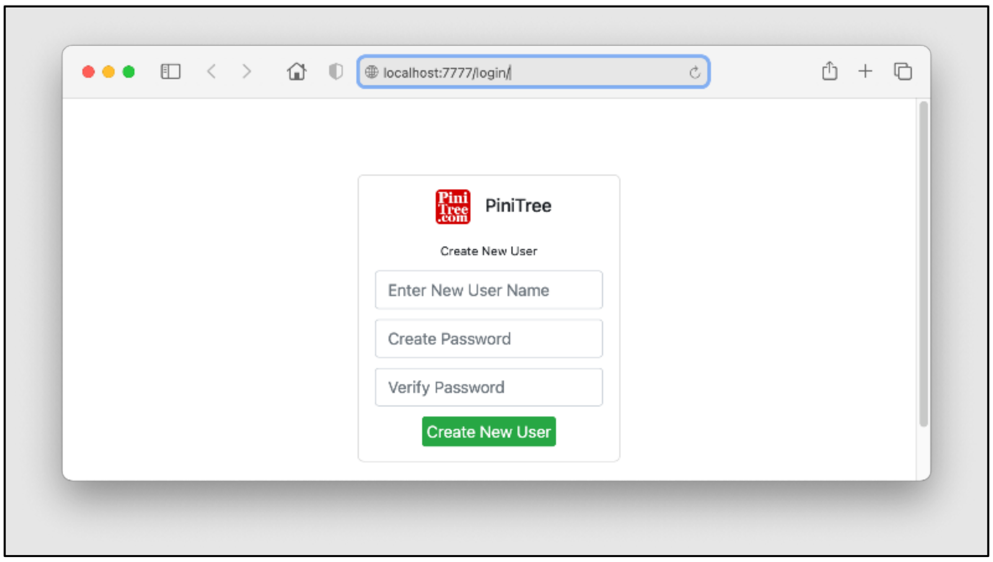


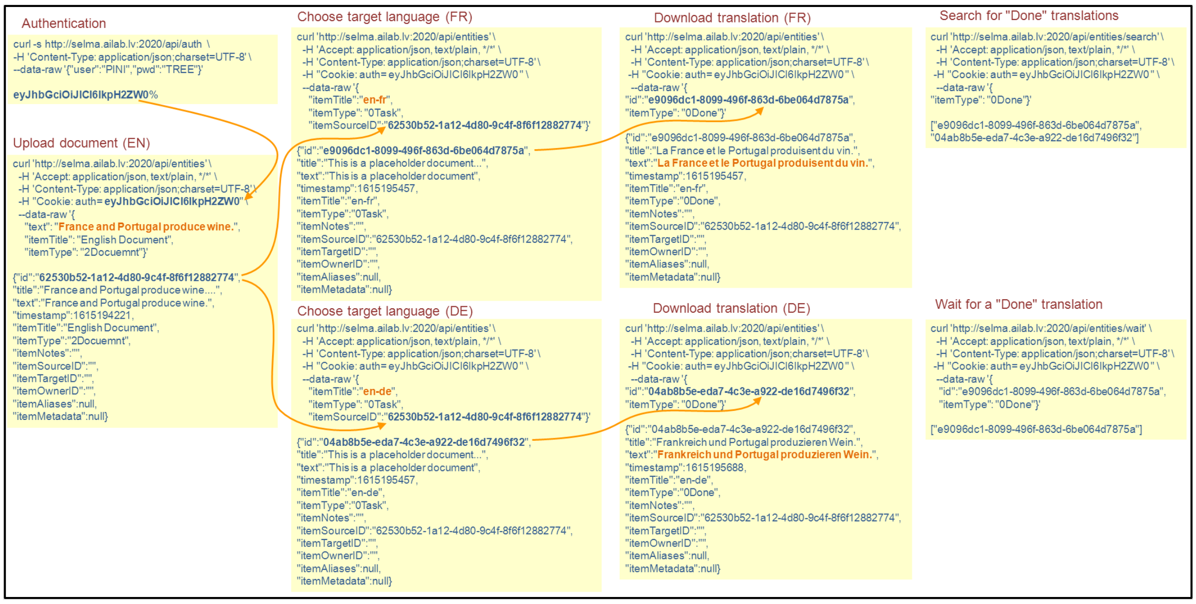
Figure 4 *Creating an admin user for the PiniTree ontology editor*

The out-of-the-box PiniTree can be used as a simple web server – click on the PiniTree.com logo in the upper left corner and upload a few files through the "Add File" button. Then click on the "Open Uploads" to browse the uploaded files. If you had also uploaded an "*index.html*" file, that would have been displayed instead of the file list. Now, anyone with the correct URL can access it.

The core use case for the PiniTree software is creating and maintaining a Named Entity Linked document store similar to Wikipedia, as described in detail in the previous sub-section.

The PiniTree server will create a */data* folder in the current directory – this is the only place where PiniTree stores all its runtime data. To backup or clone your PiniTree instance, back up or copy this */data* folder to another computer as needed. Alternatively, one can symlink the */data* folder to another disk or directory. Additionally, the PiniTree executable options can be looked up using the "*./pinitree.linux-amd64 -help*" command.

The PiniTree ontology editor can be controlled not only from the graphical user interface via a web browser, as described in Section 3.1, but also programmatically via REST API illustrated in Figure 4. A distinctive approach in this REST API is the long-polling "wait" call, which allows external systems to react in real-time to the PiniTree database changes (e.g., due to user actions or other REST API calls) without placing any control-flow logic inside the PiniTree server itself. This allows a universal PiniTree server to orchestrate multiple parallel interactive control flows simultaneously.


Authentication
```
curl -s http://selma.ailab.lv:2020/api/auth  \
-H 'Content-Type: application/json;charset=UTF-8' \
--data-raw '{"user":"PINI","pwd":"TREE"}'

eyJhbGciOiJICI6IkpH2ZW0%
```
Upload document (EN)
```
curl 'http://selma.ailab.lv:2020/api/entities' \
  -H 'Accept: application/json, text/plain, */*' \
  -H 'Content-Type: application/json;charset=UTF-8' \
  -H "Cookie: auth= eyJhbGciOiJICI6IkpH2ZW0" \
  --data-raw '{
    "text": "France and Portugal produce wine.",
    "itemTitle": "English Document",
    "itemType": "2Docuemnt"}'

{"id":"62530b52-1a12-4d80-9c4f-8f6f12882774",
"title":"France and Portugal produce wine....",
"text":"France and Portugal produce wine.",
"timestamp":1615194221,
"itemTitle":"English Document",
"itemType":"2Docuemnt",
"itemNotes":"",
"itemSourceID":"",
"itemTargetID":"",
"itemOwnerID":"",
"itemAliases":null,
"itemMetadata":null}
```
Choose target language (FR)
```
curl 'http://selma.ailab.lv:2020/api/entities' \
  -H 'Accept: application/json, text/plain, */*' \
  -H 'Content-Type: application/json;charset=UTF-8' \
  -H "Cookie: auth= eyJhbGciOiJICI6IkpH2ZW0" \
  --data-raw '{
    "itemTitle":"en-fr",
    "itemType": "0Task",
    "itemSourceID":"62530b52-1a12-4d80-9c4f-8f6f12882774"}'

{"id":"e9096dc1-8099-496f-863d-6be064d7875a",
"title":"This is a placeholder document...",
"text":"This is a placeholder document",
"timestamp":1615195457,
"itemTitle":"en-fr",
"itemType":"0Task",
"itemNotes":"",
"itemSourceID":"62530b52-1a12-4d80-9c4f-8f6f12882774",
"itemTargetID":"",
"itemOwnerID":"",
"itemAliases":null,
"itemMetadata":null}
```
Choose target language (DE)
```
curl 'http://selma.ailab.lv:2020/api/entities' \
  -H 'Accept: application/json, text/plain, */*' \
  -H 'Content-Type: application/json;charset=UTF-8' \
  -H "Cookie: auth= eyJhbGciOiJICI6IkpH2ZW0" \
  --data-raw '{
    "itemTitle":"en-de",
    "itemType": "0Task",
    "itemSourceID":"62530b52-1a12-4d80-9c4f-8f6f12882774"}'

{"id":"04ab8b5e-eda7-4c3e-a922-de16d7496f32",
"title":"This is a placeholder document...",
"text":"This is a placeholder document",
"timestamp":1615195457,
"itemTitle":"en-de",
"itemType":"0Task",
"itemNotes":"",
"itemSourceID":"62530b52-1a12-4d80-9c4f-8f6f12882774",
"itemTargetID":"",
"itemOwnerID":"",
"itemAliases":null,
"itemMetadata":null}
```
Download translation (FR)
```
curl 'http://selma.ailab.lv:2020/api/entities' \
  -H 'Accept: application/json, text/plain, */*' \
  -H 'Content-Type: application/json;charset=UTF-8' \
 -H "Cookie: auth= eyJhbGciOiJICI6IkpH2ZW0" \
  --data-raw '{
"id":"e9096dc1-8099-496f-863d-6be064d7875a",
"itemType": "0Done"}'

{"id":"e9096dc1-8099-496f-863d-6be064d7875a",
"title":"La France et le Portugal produisent du vin.",
"text":"La France et le Portugal produisent du vin.",
"timestamp":1615195457,
"itemTitle":"en-fr",
"itemType":"0Done",
"itemNotes":"",
"itemSourceID":"62530b52-1a12-4d80-9c4f-8f6f12882774",
"itemTargetID":"",
"itemOwnerID":"",
"itemAliases":null,
"itemMetadata":null}
```
Download translation (DE)
```
curl 'http://selma.ailab.lv:2020/api/entities' \
  -H 'Accept: application/json, text/plain, */*' \
  -H 'Content-Type: application/json;charset=UTF-8' \
 -H "Cookie: auth= eyJhbGciOiJICI6IkpH2ZW0" \
  --data-raw '{
"id":"04ab8b5e-eda7-4c3e-a922-de16d7496f32",
"itemType": "0Done"}'

{"id":"04ab8b5e-eda7-4c3e-a922-de16d7496f32",
"title":"Frankreich und Portugal produzieren Wein.",
"text":"Frankreich und Portugal produzieren Wein.",
"timestamp":1615195688,
"itemTitle":"en-de",
"itemType":"0Done",
"itemNotes":"",
"itemSourceID":"62530b52-1a12-4d80-9c4f-8f6f12882774",
"itemTargetID":"",
"itemOwnerID":"",
"itemAliases":null,
"itemMetadata":null}
```
Search for "Done" translations
```
curl 'http://selma.ailab.lv:2020/api/entities/search' \
  -H 'Accept: application/json, text/plain, */*' \
  -H 'Content-Type: application/json;charset=UTF-8' \
 -H "Cookie: auth= eyJhbGciOiJICI6IkpH2ZW0" \
  --data-raw '{
"itemType": "0Done"}'

["e9096dc1-8099-496f-863d-6be064d7875a",
"04ab8b5e-eda7-4c3e-a922-de16d7496f32"]
```
Wait for a "Done" translation
```
curl 'http://selma.ailab.lv:2020/api/entities/wait' \
  -H 'Accept: application/json, text/plain, */*' \
  -H 'Content-Type: application/json;charset=UTF-8' \
 -H "Cookie: auth= eyJhbGciOiJICI6IkpH2ZW0" \
  --data-raw '{
   "id":"e9096dc1-8099-496f-863d-6be064d7875a",
   "itemType": "0Done"}'

["e9096dc1-8099-496f-863d-6be064d7875a"]
```

***Figure 5** REST API access to PiniTree ontology editor database*

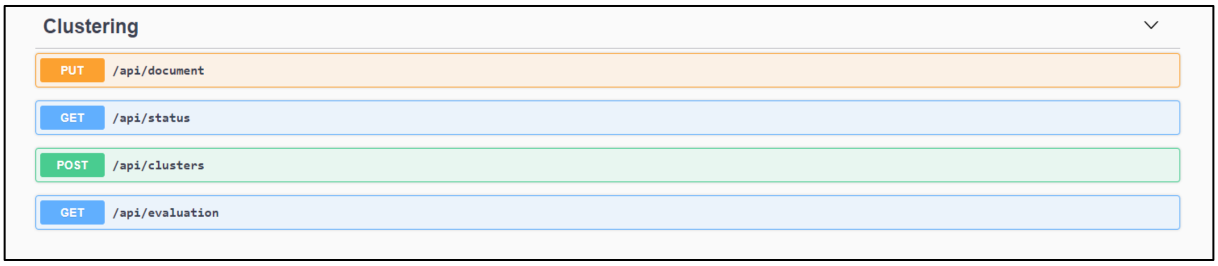
Via REST API PiniTree editor, the internal database can be integrated in real-time with the external data and processing sources and function as a component of a larger system. This is how PiniTree ontology editor is being integrated into the SELMA platform Use Case 0.

2.5 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 is underway to address this issue, although it could not be resolved within the project's timeframe. 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 (Santos et al., 2022).

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



Clustering

PUT /api/document

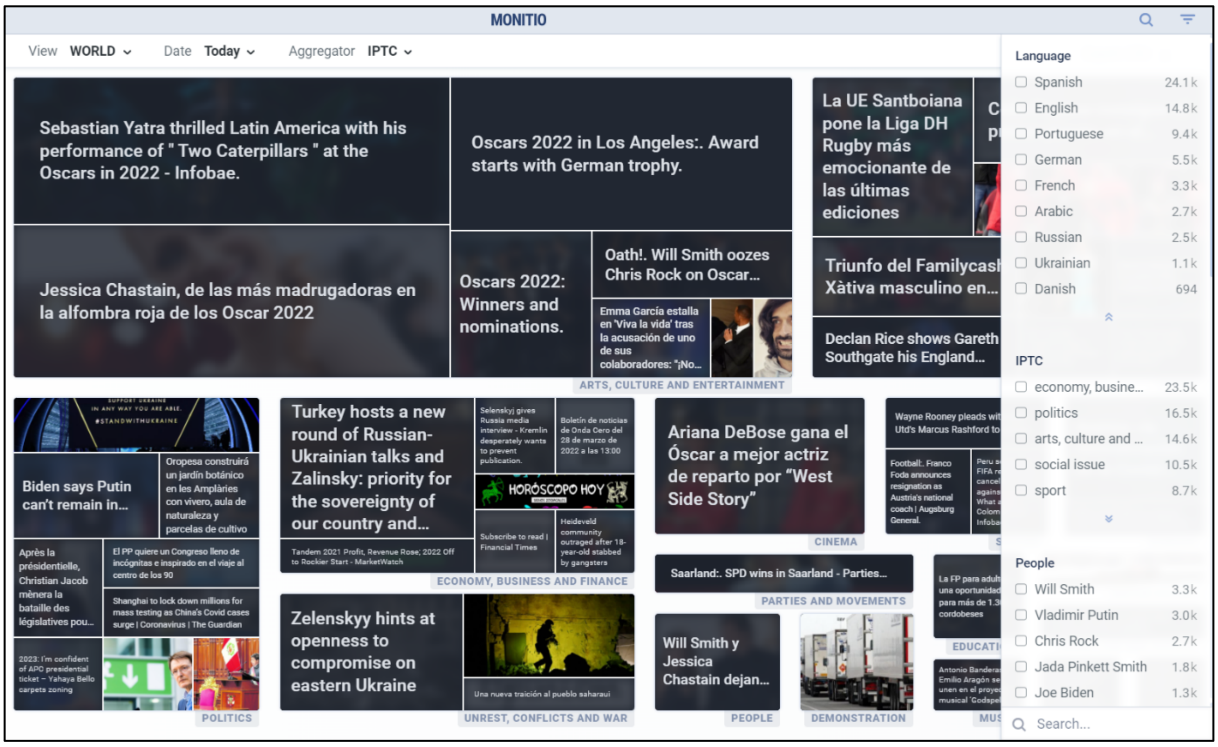
GET /api/status

POST /api/clusters

GET /api/evaluation

Figure 6 *Clustering API*

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.



MONITIO

View WORLD · Date Today · Aggregator IPTC

Sebastian Yatra thrilled Latin America with his performance of " Two Caterpillars " at the Oscars in 2022 - Infobae.

Oscars 2022 in Los Angeles:. Award starts with German trophy.

La UE Santboiana pone la Liga DH Rugby más emocionante de las últimas ediciones

Jessica Chastain, de las más madrugadoras en la alfombra roja de los Oscar 2022

Oscars 2022: Winners and nominations.

Oath!. Will Smith oozes Chris Rock on Oscar...

Emma García estalla en 'Viva la vida' tras la acusación de uno de sus colaboradores: "¡No...

Triunfo del Familycash Xàtiva masculino en...

Declan Rice shows Gareth Southgate his England...

ARTS, CULTURE AND ENTERTAINMENT

Biden says Putin can't remain in...

Oropesa construirá un jardín botánico en les Amplàries con vivero, aula de naturaleza y parcelas de cultivo

Après la présidentielle, Christian Jacob mènera la bataille des législatives pou...

El PP quiere un Congreso lleno de incógnitas e inspirado en el viaje al centro de los 90

Shanghai to lock down millions for mass testing as China's Covid cases surge | Coronavirus | The Guardian

2023: I'm confident of APC presidential ticket – Yahaya Bello carpets zoning

POLITICS

Turkey hosts a new round of Russian-Ukrainian talks and Zalinsky: priority for the sovereignty of our country and...

Selenskyj gives Russia media interview - Kremlin desperately wants to prevent publication.

Boletín de noticias de Onda Cero del 28 de marzo de 2022 a las 13:00

Tandem 2021 Profit, Revenue Rose; 2022 Off to Rockier Start - MarketWatch

Subscribe to read | Financial Times

Heideveld community outraged after 18-year-old stabbed by gangsters

ECONOMY, BUSINESS AND FINANCE

Zelenskyy hints at openness to compromise on eastern Ukraine

Una nueva traición al pueblo saharaui

UNREST, CONFLICTS AND WAR

Ariana DeBose gana el Óscar a mejor actriz de reparto por "West Side Story"

CINEMA

Wayne Rooney pleads wit Utd's Marcus Rashford to

Football: Franco Foda announces resignation as Austria's national coach | Augsburg General.

Saarland:. SPD wins in Saarland - Parties...

PARTIES AND MOVEMENTS

Will Smith y Jessica Chastain dejan...

PEOPLE

DEMONSTRATION

EDUCATI

MUS

Language

Spanish	24.1k
English	14.8k
Portuguese	9.4k
German	5.5k
French	3.3k
Arabic	2.7k
Russian	2.5k
Ukrainian	1.1k
Danish	694

IPTC

economy, busine...	23.5k
politics	16.5k
arts, culture and ...	14.6k
social issue	10.5k
sport	8.7k

People

Will Smith	3.3k
Vladimir Putin	3.0k
Chris Rock	2.7k
Jada Pinkett Smith	1.8k
Joe Biden	1.3k

Search...

Figure 7 *"Storylines" Dashboard from Monitio Showing Multilingual Clustering*

During the second reporting period this component had been updated in order to improve its performance. Namely, the communication pipeline has been changed to use gRPC (a high-performance Remote Procedure Call) instead of REST and the embeddings creation has been decoupled from the module.

2.6 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.

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 good5. 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.

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 unexpected 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. The linguist team at Priberam evaluated a set of 666 documents in several languages comparing whether the set of topics attributed using the old model was better or worse than the last version and the result was that 214 documents had a better set of labels using the old and 436 where better using the latter, and 16 was indifferent. The next table shows the evaluation for each language.

Language	Indifferent	Old is better	New is better
ar	2	17	24
da		18	32
de	1	13	30
el		3	22
en	1	20	31
es		12	34
fa		8	33
fr		16	34
hu	2	17	30
it	1	12	33
ja			1
lv		13	27

pt	2	17	22
ru	5	15	27
tr	2	20	21
uk		13	35
	16	**214**	**436**

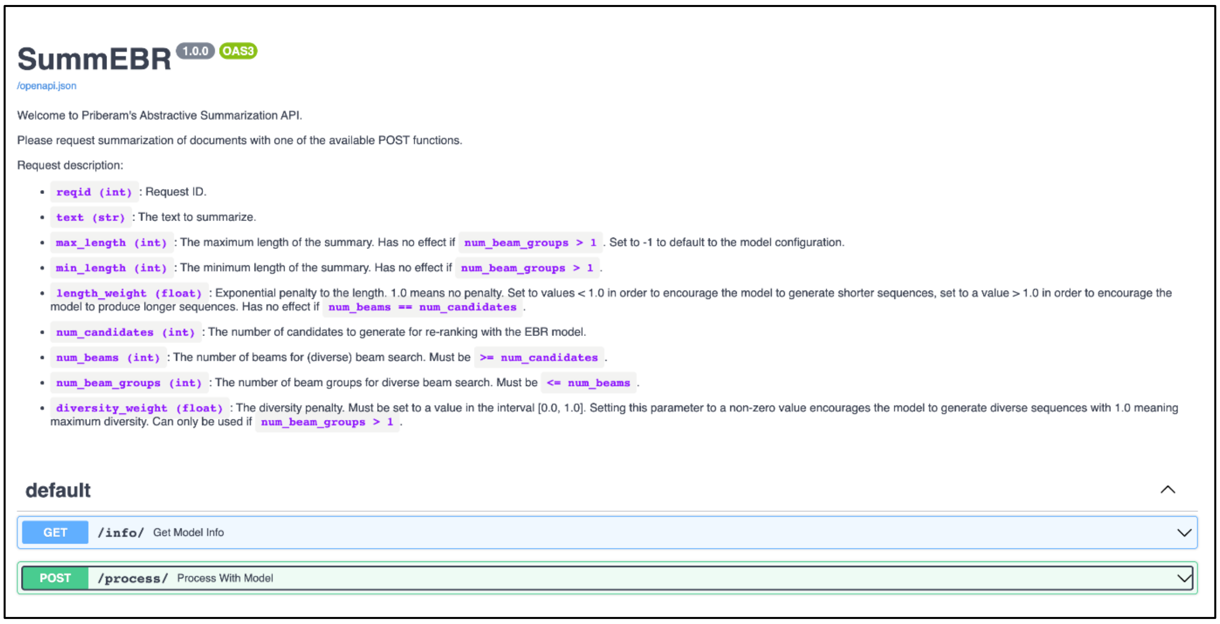
Table 3 *User topic detection evaluation*

2.7 News Summarization

Monolingual Abstractive 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. The former was approached with a re-ranking approach as described in Pernes et al. (2022) and reported in D2.7.

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. We will also release a summarization component for cross-lingual summarization as soon as our research on this problem is concluded.



SummEBR 1.0.0 OAS3

/openapi.json

Welcome to Priberam's Abstractive Summarization API.

Please request summarization of documents with one of the available POST functions.

Request description:

- `reqid (int)` : Request ID.
- `text (str)` : The text to summarize.
- `max_length (int)` : The maximum length of the summary. Has no effect if `num_beam_groups > 1` . Set to -1 to default to the model configuration.
- `min_length (int)` : The minimum length of the summary. Has no effect if `num_beam_groups > 1` .
- `length_weight (float)` : Exponential penalty to the length. 1.0 means no penalty. Set to values < 1.0 in order to encourage the model to generate shorter sequences, set to a value > 1.0 in order to encourage the model to produce longer sequences. Has no effect if `num_beams == num_candidates` .
- `num_candidates (int)` : The number of candidates to generate for re-ranking with the EBR model.
- `num_beams (int)` : The number of beams for (diverse) beam search. Must be `>= num_candidates` .
- `num_beam_groups (int)` : The number of beam groups for diverse beam search. Must be `<= num_beams` .
- `diversity_weight (float)` : The diversity penalty. Must be set to a value in the interval [0.0, 1.0]. Setting this parameter to a non-zero value encourages the model to generate diverse sequences with 1.0 meaning maximum diversity. Can only be used if `num_beam_groups > 1` .

default

GET `/info/` Get Model Info

POST `/process/` Process With Model

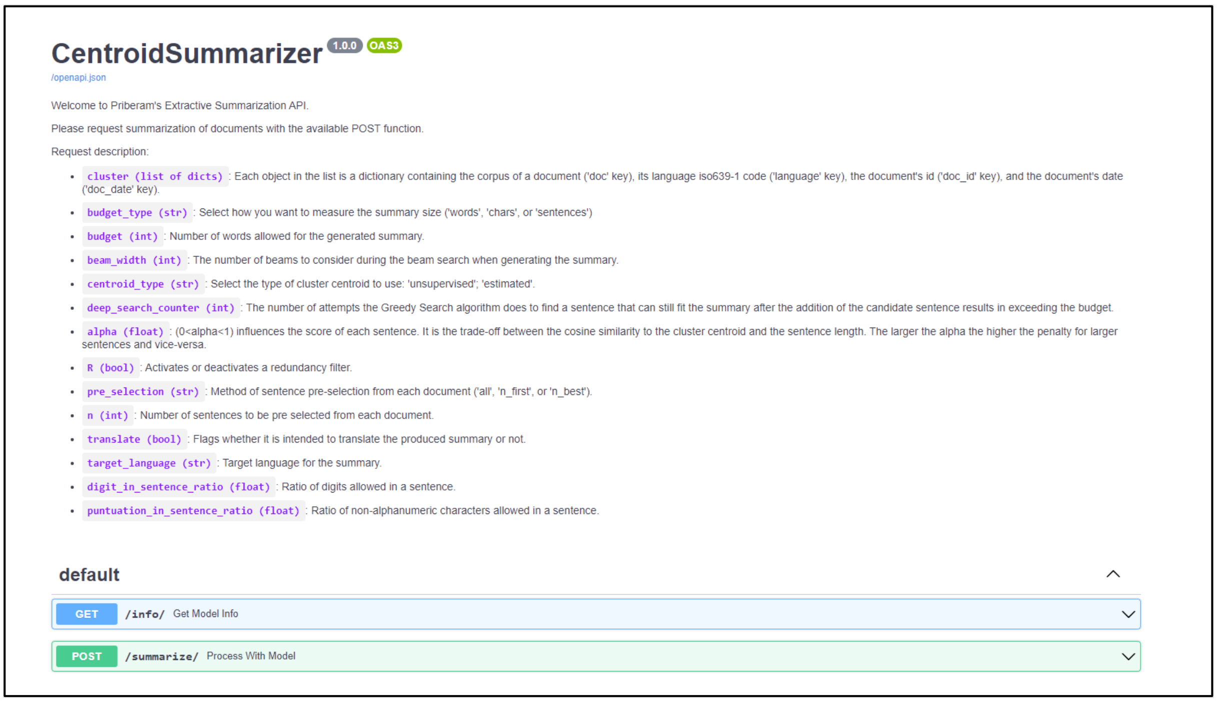
Figure 8 *Abstractive Summarization API*

Multilingual Multi-Document Extractive Summarization

In addition to our ongoing research in abstractive multilingual and cross-lingual summarization, we have also concentrated on devising an efficient method for extracting the most pertinent segments from a multilingual cluster of documents. This issue was tackled as outlined in Gonçalves et al. (2023) and documented in D2.7.

The Priberam's Extractive Summarization API interacts with our centroid-based summarization algorithm, which comprises three submodules: a multilingual sentence encoder (Reimers & Gurevych., 2020), the centroid estimation model described in Gonçalves et al., 2023, and a machine translation model (M2M-100, Fan et al., 2020) that translates the summary to the desired target language.

The proposed methodology has been implemented as an API integrated into Priberam's product, Monitio, offering extractive summarization capabilities for multilingual clusters of news articles. The underlying model was trained on English, French, and Spanish documents but was evaluated showing robust performance in other languages as well, as reported in D2.7. This is due to the multilingual nature of the sentence embeddings provided by *distiluse-base-multilingual-cased-v2*, which supports more than 50 languages. The API interface is illustrated in Figure 4.



CentroidSummarizer 1.0.0 OAS3

/openapi.json

Welcome to Priberam's Extractive Summarization API.

Please request summarization of documents with the available POST function.

Request description:

- `cluster (list of dicts)` : Each object in the list is a dictionary containing the corpus of a document ('doc' key), its language iso639-1 code ('language' key), the document's id ('doc_id' key), and the document's date ('doc_date' key).
- `budget_type (str)` : Select how you want to measure the summary size ('words', 'chars', or 'sentences')
- `budget (int)` : Number of words allowed for the generated summary.
- `beam_width (int)` : The number of beams to consider during the beam search when generating the summary.
- `centroid_type (str)` : Select the type of cluster centroid to use: 'unsupervised'; 'estimated'.
- `deep_search_counter (int)` : The number of attempts the Greedy Search algorithm does to find a sentence that can still fit the summary after the addition of the candidate sentence results in exceeding the budget.
- `alpha (float)` : (0<alpha<1) influences the score of each sentence. It is the trade-off between the cosine similarity to the cluster centroid and the sentence length. The larger the alpha the higher the penalty for larger sentences and vice-versa.
- `R (bool)` : Activates or deactivates a redundancy filter.
- `pre_selection (str)` : Method of sentence pre-selection from each document ('all', 'n_first', or 'n_best').
- `n (int)` : Number of sentences to be pre selected from each document.
- `translate (bool)` : Flags whether it is intended to translate the produced summary or not.
- `target_language (str)` : Target language for the summary.
- `digit_in_sentence_ratio (float)` : Ratio of digits allowed in a sentence.
- `puntuation_in_sentence_ratio (float)` : Ratio of non-alphanumeric characters allowed in a sentence.

default
GET /info/ Get Model Info
POST /summarize/ Process With Model

Figure 9 *Multilingual, multi-document summarization API*

2.8 Story Segmentation

Speaker diarization is a cutting-edge process in audio processing, which involves partitioning an audio stream into homogenous segments according to the speaker's identity. It is often colloquially referred to as "who spoke when." The ability to accurately label multiple speakers enhances a voice-driven application's effectiveness and contributes significantly to the analytical assessment of audio recordings.

Our final release is built upon advanced machine learning algorithms developed on a vast dataset of diverse speech samples from the VoxCeleb dataset:

https://www.robots.ox.ac.uk/~vgg/data/voxceleb, which is extracted from celebrity interview videos on YouTube containing speeches from over 7,000 speakers. It analyses unique voice features, such as pitch, tone, and speech patterns, and then utilizes these characteristics to distinguish between speakers. The flowchart provided below illustrates the three-step process involved in preparing a speaker segmentation system:

- Step 1 — Voice Activity Detection: The initial step in the diarization process involves Voice Activity Detection (VAD), which identifies speech segments within the audio stream. The VAD component effectively differentiates between speech and non-speech intervals, ensuring that subsequent steps focus only on spoken content.
- Step 2 — Speaker Embedding Extraction: Following VAD, the system extracts speaker embeddings. These embeddings are high-dimensional vectors that represent the unique characteristics of each speaker's voice. This process leverages the distinctive features of vocal patterns to differentiate between speakers.
- Step 3 — Speaker Clustering: The final step is clustering, which involves grouping the extracted embeddings into clusters, each representing a different speaker. This step is crucial for attributing the identified speech segments to individual speakers, resulting in the final output of clearly labeled speaker segments.

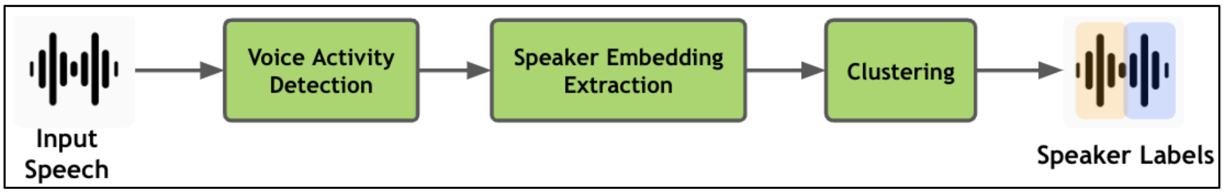
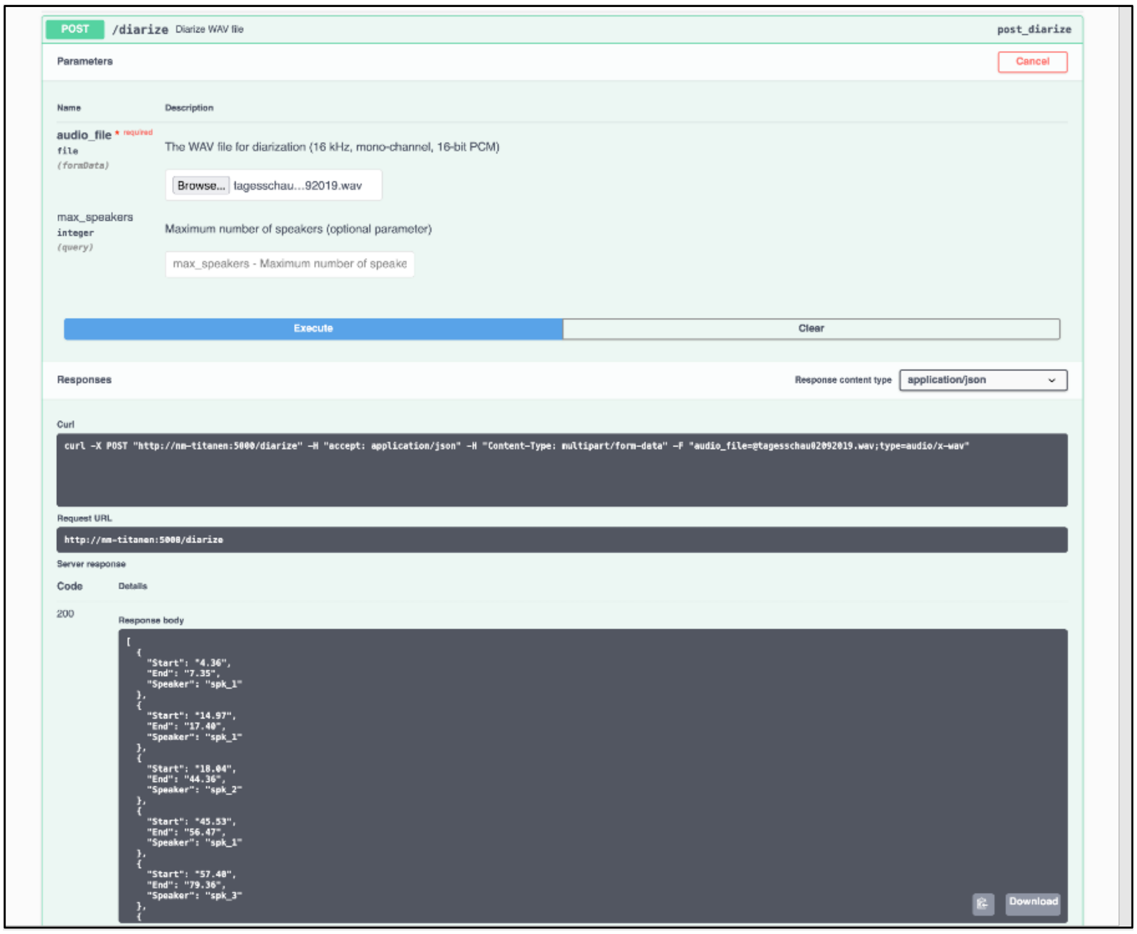


Figure 9 *Speaker Segmentation Flowchart of Streamlined and Three-step Process for Diarization*

A key advantage of the final release is its optimization of CPU usage. Each step in the workflow, from voice activity detection to speaker clustering, has been tuned to run efficiently on CPUs. This optimization eliminates the need for expensive GPU resources, ensuring that our tool is accessible and cost-effective for users who may not have access to specialized hardware.

To enhance accessibility and ease of integration, our system is encapsulated within a web service accessible via a RESTful API. The screenshot below showcases the Swagger API interface, a user-friendly framework for interacting with the API. Its endpoint accepts an audio file and returns the speaker-segmented data. It requires the audio file in 16 kHz, mono-channel, 16-bit PCM, and an optional parameter to indicate the maximum number of speakers expected in the audio. The response is returned in JSON format, which includes the start time, end time, and speaker label for each identified segment.

Our segmentation tool is engineered for high efficiency and speed. It can process 15 minutes of audio in less than two minutes, offering a swift turnaround suitable for quick analysis of large datasets. The efficiency of the tool is further enhanced by extracting speaker embeddings through the ONNX (Open Neural Network Exchange) file format. ONNX provides a more accelerated execution of models compared to the conventional Torch binaries, contributing significantly to the tool's rapid processing capabilities. This optimized model format ensures faster computations, particularly when operating on CPU environments.



/diarize Diarize WAV file post_diarize

Parameters

Name	Description
audio_file * required `file` *(formData)*	The WAV file for diarization (16 kHz, mono-channel, 16-bit PCM) tagesschau…92019.wav
max_speakers `integer` *(query)*	Maximum number of speakers (optional parameter) max_speakers - Maximum number of speake

Responses

Response content type: application/json

Curl

```
curl -X POST "http://nm-titanen:5000/diarize" -H "accept: application/json" -H "Content-Type: multipart/form-data" -F "audio_file=@tagesschau02092019.wav;type=audio/x-wav"
```

Request URL

```
http://nm-titanen:5000/diarize
```

Server response

Code	Details
200	Response body

```
[
  {
    "Start": "4.36",
    "End": "7.35",
    "Speaker": "spk_1"
  },
  {
    "Start": "14.97",
    "End": "17.40",
    "Speaker": "spk_1"
  },
  {
    "Start": "18.04",
    "End": "44.36",
    "Speaker": "spk_2"
  },
  {
    "Start": "45.53",
    "End": "56.47",
    "Speaker": "spk_1"
  },
  {
    "Start": "57.40",
    "End": "79.36",
    "Speaker": "spk_3"
  },
  {
```

Figure 10 *Speaker Diarization API Interface using the Swagger UI*

It is important to note that while the tool is optimized for speed, the complexity of speaker clustering increases exponentially with longer audio inputs. Therefore, we recommend using the tool for audio files shorter than one hour to maintain optimal performance. For audio inputs that exceed this recommended duration, the clustering phase may require additional processing time due to the increased computational load.

3. Conclusions

The culmination of our efforts in SELMA has resulted in a suite of software tools targeting various tasks outlined in WP2, spanning T2.1 through T2.5. Notable among these releases is our entity linking software, which boasts extensive language support and has demonstrated exceptional performance on various test datasets. In addition, our efforts in language transfer for named entity recognition have yielded impressive results. Furthermore, the successful integration of multilingual news classification and summarization modules represents a significant step forward. Our efforts have also resulted in significant enhancements to the language support of our models, providing improved scalability and performance in production environments. In addition, the introduction of the first language-agnostic speaker segmentation module is a pioneering achievement in our field. Overall, these software releases underscore our commitment to advancing research and innovation in multilingual natural language processing.

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.*

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., ... & Joulin, A. (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research,* 22(107), 1-48.

Gonçalves, S., Correia, G., Pernes, D., & Mendes, A. (2023). Supervising the Centroid Baseline for Extractive Multi-Document Summarization. *Proceedings of the 4th New Frontiers in Summarization Workshop*, Singapore. Association for Computational Linguistics (pp. 87-96).

Paikens, P., Barzdins, G., Mendes, A., Ferreira, D., Broscheit, S., Almeida, M., Miranda, S., Nogueira, D., Balage, P., & Martins, A. (2016a). SUMMA at TAC Knowledge Base Population Task 2016, *DOI: 10.5281/zenodo.827317*

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).*

Reimers, N., & Gurevych, I. (2020, November). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4512-4525).

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*.

Znotiņš, A. & Barzdins, G. (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.*