Research and Innovation Action (RIA) H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu

D2.7 Final Report on Continuous Massive Stream Learning

Work Package	2
Main Author	Diogo Pernes
Co-Authors	Afonso Mendes, Gonçalo Correia, João Figueira, João Santos, Tugtekin Turan
Reviewer	Yannick Estève
Version	1.0
Contractual Date	March 31, 2024
Delivery Date	March 28, 2024
Dissemination Level	Public

Version History

Version	Date	Description
0.1	01/02/2024	Initial Table of Contents (ToC)
0.2	01/02/2024	Initial Input
0.3	14/03/2024	Internal Review
1.0	19/03/2024	Publishable version

Executive Summary

This report summarizes the scientific progress of our natural language tasks in work package two (WP2) of the SELMA project. The goal of WP2 is to enable the SELMA system to automatically learn from a large live multilingual data stream. In this document, we report the progress of our work on each of the subtasks required to achieve the WP2 goals: cross-lingual stream representations, named entity recognition and linking, story segmentation, news classification, clustering, and summarization. In the first two sections, we present the general framework and overview of WP2, particularly introducing each subtask separately. We then define our methods for each task in Section 3 and present our experimental results in Section 4. The concluding section of this report summarizes the significance of our findings in the context of SELMA, and also suggests avenues for future research efforts.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

Continuous learning aims to enable information systems to learn from a continuous data stream across time. We, as human beings, can learn by building on our memories and applying past knowledge to understand new concepts. However, it is not easy for existing deep learning architectures to learn a new task without forgetting previously acquired knowledge. Unlike humans, existing machine learning ideas are primarily trained in an isolated environment and can be used effectively only for a limited time. Therefore, the produced models become less accurate over time due to the changing distribution or nature of the data. With the recent advancements in deep learning, the problem of continuous learning in natural language is becoming even more critical, as current approaches cannot effectively keep previously learned knowledge and adapt to new information simultaneously.

The SELMA continuous learning platform specifically targets multilingual broadcast monitoring and production. With the exponential growth of online news content in several languages, the challenge is to avoid a language and cultural bottleneck. Hence, this work package eventually brings together many sources and makes information accessible to users in multiple languages yet keeping relevant knowledge present in the original multilingual data sources.

Multilingualism supports the opportunity of sharing valuable knowledge across languages. We, therefore, aim to propose a unified approach to multilingual media monitoring and content production by contributing to recent advances in deep learning, particularly breakthroughs in knowledge and language transfer and fine-tuning of task models from user feedback. High-quality and up-to-date cross-lingual text and entity representations are vital components of this work package. Computing and updating these representations via user feedback is an important research direction in the context of natural language on news data, as relevant entities, which have a defining role in news stories, take part in ever-evolving story contexts.

To this end, this work package presents the research results leading to a high-performance modular platform for ingesting and processing data streams with the goal of training and maintaining multilingual natural language components. Our proposed methods create a unique framework for integrating high-quality user feedback with massive amounts of multilingual data. Low-resource

languages have also been addressed due to the multilingual data context combined with transfer learning approaches.

2. Architecture

This work package enables the SELMA platform which ingests a large-volume live multilingual stream of documents and continuously incorporates knowledge to update the models. Moreover, transfer learning was adopted to improve model performance on low-resource languages with knowledge from high-resource languages.

The multilingual stream is at the core of the SELMA processing pipeline. A collection of news sources serves as a reference to guide the natural language downstream tasks executed on the user-supplied data. We mainly researched novel approaches to jointly extract named entities from the reference stream and link them to a knowledge base to enable the proposed methods. We also employ current practices to learn up-to-date contextual cross-lingual embedding representations for text/entities and efficiently search on these representations.

In summary, the main achievements of this work package are:

- Learning a multilingual representation for text and entities from Wikipedia, Wikidata and the input reference stream
- Detecting mentions to named entities in multilingual scenario
- Identifying named entities and linking them to a knowledge base
- Incorporating the user feedback into training and improvement of our models
- Transferring knowledge between languages, to the benefit of low-resourced languages

To achieve these goals, we can define the primary tasks of this work package as follows:

Cross-Lingual Stream Representations

This task focuses on learning contextual word and entity representations captured from a live news article stream. Note that the extensive data scale makes this task particularly challenging, in addition to the emphasis on serving across several languages simultaneously. Hence, to enable knowledge transfer from higher- to lower-resourced languages, we learned a cross-lingual representation space, i.e., a representation where word contexts from different languages are mapped into a shared space, to enable knowledge transfer from higher- to lower-resourced languages.

Named Entity Recognition and Linking

The purpose of this task was to develop statistical models for detecting entities within news article streams and mapping these entities to a knowledge base link. This step is fundamental to perform content enrichment on the stream. Therefore, we focused on deep contextualized entity representations, where we first detect entity mentions and then perform entity disambiguation to obtain the correct link to the knowledge base. Our approach achieves state of the art results for entity linking in English documents and shows very robust performance in zero-shot transfer to less-resourced languages. An important consideration in our approach was the identification of nil entities, referring to entities lacking corresponding entries in the knowledge base.

Story Segmentation

This task aims to segment long audio segments into meaningful units, providing speaker clustering, speaker recognition, and topic segmentation. For speaker clustering, the identity of the speakers is unknown, and the system provides only labels for segments of the same speaker appearing multiple times in one file. This is useful, e.g., for interviews in which only the statements of a particular user are of interest to the journalist. Moreover, the human voice can contain personal attributes of unique pronunciation (vocal tract shape) and speaking manner (accent and rhythm). Therefore, speaker recognition is defined as the task of identifying people from their voices. We approached this problem with an end-to-end framework for the recognition of specific speakers from a known speaker database. On the other hand, we investigated the speaker diarization task to label news content with classes that correspond to speaker identity to address "who spoke when".

Online News Classification and Clustering

News classification aims at categorizing a given text sequence with one (or more) predefined class label(s) describing its semantic content. To this end, we followed recent research on cross-lingual representations for topic labeling across different languages using deep contextual models. One of our main concerns was to obtain a common space for different label sets on multilingual data. This cross-lingual space was then used to automatically cluster news documents from different languages related to the same story. The developed clustering algorithm can work in an online environment where the document stream is processed, and documents are clustered in real time.

News Summarization

This task focuses on summarizing news content using state-of-the-art abstractive neural approaches. The biggest challenge in this task has been and still is the presence of factual inconsistencies in the generated summary. Both automatic detection and mitigation of factual inconsistencies are still open research problems. In our research, we have taken steps towards mitigating these problems using a re-ranking model that successfully selects the best summary from a collection of candidates. In addition, and given the multilingual nature of SELMA, we considered the problem where we want to produce summaries of a document in multiple target languages, while ensuring semantic consistency across target languages. For this task, we compare a re-ranking approach based on cross-lingual embedding similarity with the usual pipeline of summarization and translation models. Finally, we also addressed the multilingual, multi-document setting using a novel extractive approach that estimates a contextual multilingual representation for a cluster of documents prior to the sentence selection step.

In a parallel line of work, we also addressed the challenging problem of end-to-end speech-to-text summarization. This task has rarely been explored in previous work, which usually tackles the problem using a cascade of transcription and text summarization modules. The rationale behind an end-to-end approach is to avoid error propagation and to open the possibility of increased computational efficiency. The approach we propose yields promising results, although it is still inferior to the cascade of transcription and summarization.

3. Scientific Approach

This section presents an introduction to our proposed methodologies employed for this deliverable. The problems presented in the introduction section will be explained in detail under the following sub-chapters. We will give experimental results and their discussions in the next section.

3.1. Named Entity Recognition

For the Named Entity Recognition (NER) task, we investigate two ideas: Hierarchical Nested Named Entity Recognition (HNNER) and example-based NER. In the following subsections, we present a summary of these approaches. During the second reporting period we investigated the behavior of the proposed models in a multilingual scenario and their ability to zero-shot to unseen languages during training.

In the last period we investigated a new model architecture, derived from the HNNER approach, with better computing performance and a smaller number of parameters using an attention only approach. We also investigated the possibility of training jointly several datasets annotated with different ontologies with a two-fold objective: increase performance leveraging on cross-ontology transfer, and the ability two deploy a single production model for several ontologies.

Hierarchical Nested NER

The task of recognizing mentions to entities in text can take different forms. We focus on the hierarchical nested approach, as shown in Figure 1, where a given sequence of words can correspond to more than one entity type, e.g., “gpe” and “gpe → city”, with “city” being a more fine-grained entity type, with the added possibility of including entities within entities (nested entities) as shown in the example bellow. This subsection reports two approaches related to the task of hierarchical nested NER: improvements made to Marinho et al. (2019) (Stack-LSTM), and a new biaffine approach, heavily based upon Yu et al. (2020).

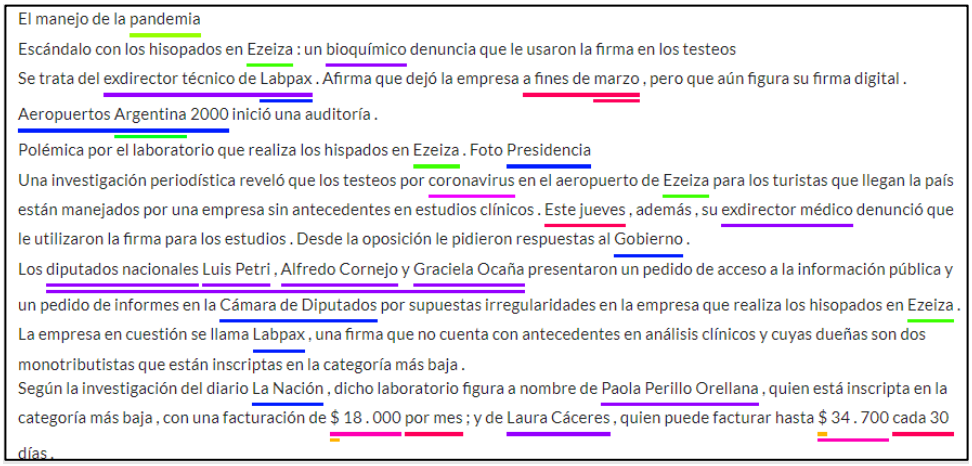

Figure 1 *Nested NER annotation example*

The Stack-LSTM approach models hierarchical and nested entities via four main actions: transitions, shifts, reduction, and outs. These actions modify the system's state by interacting with the words in an input sentence over a series of "stacks", which model different aspects using LSTMs. All words are represented by concatenating their corresponding fixed-word lookup embedding and learned character sequence embedding representations. We propose replacing the original word representations by contextual embedding representations, using existing models based on architectures such as BERT (Devlin et al. (2019)), coupled with an extensive study of pooling approaches and fine-tuning strategies. The main advantage is to use more powerful pre-trained embedding models, which can leverage the context of a word within its sentence. Several works highlight the excellent performance of applying pre-trained multilingual contextual embedding to languages other than English.

The Biaffine model follows the work of Yu et al. (2020). This model scores pairs of start and end tokens in a sentence to explore all spans so that the model can predict named entities accurately. We propose using a biaffine classifier model, initially capable of identifying flat and nested entities. It uses token-level representations based on a combination of character and pre-trained contextual embeddings coupled with a biaffine model. This returns a score tensor of every possible class of start-end span combinations. It has dimensions n x n x c, where n is the number of tokens in the input, and c is the number of classes plus one, the no-entity class. We introduce three changes to make this approach capable of modeling hierarchical entities: (i) the score tensor, which is an output of the

biaffine model, is now *n* x *n* x *m*, where *n* corresponds to the number of tokens in the input, and *m* corresponds to a span embedding dimension; (ii) we add a classifier that predicts whether a span corresponds to an entity or not. The intuition is that since predicting multiple labels for each span will involve evaluating all possible spans sequentially, skipping as many spans as possible improves performance; (iii) using the score tensor, we use an LSTM model to predict entities for a given span at a time, until the "end of the sentence" token is predicted. At each step, the LSTM model input becomes the concatenation of different intermediate representations.

Example-Based NER

Current research in text generation has shown that combining a traditional generation model with a k-nearest neighbors (kNN) approach improves performance (Khandelwal et al. (2020), Khandelwal et al. (2021)). We explore the possibility of extending these approaches to the NER task. In particular, for each token of the input sentence, we find the closest *k* tokens on a set of similar sentences retrieved using sentence embeddings (SBERT) (Reimers et al. (2019)). Then, we follow either a single-*k* approach, where the kNN distribution for each token is obtained from a single *k* value, or a multi-*k* approach, where the kNN distribution for each token is the average of the distributions obtained for multiple *k* values. The remaining steps follow the works mentioned above.

We highlight the possibility of using this approach to leverage user feedback by continuously adapting the NER predictions with the data collected, avoiding re-training the model as often. We aim to use this approach to deal with user feedback for entity linking and the NER from speech.

Learning Cross-language NER

In SELMA, one of our main objectives is to obtain good models over a large number of languages. Additionally, one of our major concerns is scalability when moving the models to production, which means that we cannot deploy one model for each of the SELMA target languages. To attain these objectives, we researched the possibility of having only one model for all the target languages without losing performance. We also wanted to know if it is possible to improve each of the languages by using data from other languages.

To meet the above objectives, we researched the following approaches keeping in mind that our aim, to keep platform-needed resources to a minimum, is to have one model that covers all languages:

- Training one joint model using as training data the mixture of all language datasets annotated with the common ontology making no explicit difference between them.

- Introduce an additional language class, and their respective *transition and reductions,* on the stack model representing the language of the input document.

- Use of language adapters that can be trained and injected in the model for each language as proposed by Neil Houlsby et al. (2019).

In the evaluation section, we report the results of the first hypothesis above, which proved very good and more general than the remaining.

Learning Cross-Ontology NER

In MONITIO (UC2) we need to detect mentions of different types, like the most common ones (persons, organizations, locations, etc..) for which we have annotated a multilingual dataset, and other kinds of mentions like medicine ones (genes, diseases, proteins, etc..) or legal. Available to the community there are quite a few NER datasets specialized on different domains and using different ontologies, our research in focused on extending our NER model so that it could learn simultaneously from different datasets and simultaneously improve the performance on each of them. The additional advantage of such a model would be to on one single pass process documents and output NER mentions in multiple ontology thus saving a lot of production expensive resources like GPUs.

To this end we will propose a new architecture based on the Stack-LSTM transition-based model where we replace the LSTM by an attention mechanism like the one used by Ganea and Hoffmann (2017) on Entity Liking where we model the transitions as a rolling attention over the actions as defined by us in Marinho et al. (2019). At the time of writing the experimental work is still ongoing but preliminary results show and increase from 0.77 to 0.79 F1 on the GENIA dataset (Kim, Jin-Dong & Ohta et. al.), a semantically annotated corpus for bio-textmining, when comparing our Stack-LSTM to our new model.

We also propose a new method to train a multi-ontology/multi-dataset model by during training allowing to modes of operation: one where the model acts as a teacher when it predicts a tag from a different ontology then the ones of the current dataset and another where the model follows the gold

from the current dataset. Preliminary tests show that this approach improves the results in most of the datasets tested. We expect to publish this work shortly after the completion of the SELMA project.

3.2. Entity Linking and Cross-Lingual Stream Representations

Entity linking (EL) is the task of connecting a named entity in a document to an entry in a Knowledge Base (KB). One way to address this problem is to create a candidate set for each named entity with possible entities from the KB and then rank the candidates to choose the most likely entity to be linked. Our work follows this approach and employs a model inspired by Dynamic Context Augmentation (DCA) by Yang et al. (2019), which is an improvement over the original model proposed by Ganea and Hoffmann (2017). This family of models have two main components: the pre-trained entity embeddings and the ranking model (based on the DCA) that uses those embeddings and scores candidates through a combination of independent scores. This formulation allows for an existing subset of entities to be adapted or added without retraining the whole set of entities in the knowledge base, facilitating user feedback and stream learning scenarios.

Our entity embeddings are bootstrapped from a frozen set of contextual word embeddings. Following the idea in Yang et al. (2019), we employ Wikipedia hyperlinks as the context to create embeddings from. The original articles leveraged English Word2Vec (Mikolov et al. (2013)) embeddings. However, we further extended the capability of our entity embeddings by leveraging the contextual representations of XLM-RoBERTa (Conneau et al. (2020)). This is a transformer model trained in 100 different languages, which obtains state-of-the-art results in many cross-lingual benchmarks. We further improve the pipeline by using a finetuned version of XLM-RoBERTa (XLM-R) that was trained along with our HNNER approach, bringing us closer to an end-to-end approach to NER and EL. The reasons for this change in the backbone model behind the entity embeddings was then three-fold: (1) to unite through a common model NER and EL; (2) to have a backbone model that not only has an extensive set of languages available but also outperforms others in cross-lingual tasks; and (3) to create entity embeddings from contextual representations, which is known to bring improvements.

The ranking model DCA receives a pre-computed candidate set for each mention and yields a score for each candidate, choosing the highest scoring candidate as the linked entity. This score is a composition of independent scores. Yang et al. (2019) model considered five scores: (i) prior

probability of the Entity given the mention, *P(E|m) where E is the entity and m the mention*, computed using Wikipedia hyperlink count frequency; (ii) a local disambiguation score that calculates an attention score between a candidate embedding and word embeddings surrounding the mention to assign higher importance to certain context words; (iii) a global entity coherence score to produce an attention score between a candidate and previously disambiguated entities, under the assumption that there is consistency between document mentions; (iv) a score based on classifying the mention as four entity types (PER, GPE, ORG, UNK); and (v) a score based on the closely associated entities of previously disambiguated entities. We modified this model by considering all mentions in the document for the global score and by adding a more complex score on the mentioned type of classification. The former considers the top candidates of all mentions in the document, based on the local score given by (ii); this greatly improves the training and inference speed of the model, as it is no longer necessary to iteratively compute this global score as more entities are disambiguated. The latter generates the cosine similarity between the predicted type embedding of a mention and the type embedding of a candidate. The mentioned type is inferred using a classifier following Cardoso et al. (2020), trained alongside DCA and uses more than 40 mention types.

During the first reporting period, our learned embeddings vocabulary can consider exclusively English, German, and Portuguese Wikipedia pages. In a multilingual scenario, for a given entity, we would sample positive words from the English Wikipedia page and hyperlinks if that entity has an English Wikipedia page. Otherwise, we would sample positive words from the respective language from which it was obtained, either German or Portuguese. To train the DCA model, we used different configurations: training on the English portion of the CoNLL-2003 (Tjong Kim Sang & De Meulder (2003)) NER shared task data, containing news stories from Reuters news agency; training on an English Wikipedia page set where hyperlinks are considered mentions and their linked pages are the gold entities; training on both sets simultaneously. The latter experiments did not reach better results, so we ended up only using CoNLL-2003 as the training set.

During the second reporting period we started investigating how the previous approaches behaved when we extended the number of languages to a much bigger set of languages (currently 40). We found out that the models behave similarly when training the entity representations using these additional languages. Our objective was to have a single model for EL and NER trained jointly or at least sharing the same base contextual model, which we achieved by sharing the same base model of

XLM-R. For that purpose, we started researching the possibility of using the base contextual model fine-tuned on the NER data to learn the entity embeddings. This implies changing the Ganea model presented above to gather negative token embeddings from the Wikipedia multilingual dataset, posing a big optimization challenge due to the size of the dataset.

To train the contextual entity embeddings, we defined the following procedure, based on the procedure from Ganea and Hoffmann (2017):

1. Initialize the entity embeddings with the mean pooling of the individual token embedding of the entity title.
2. Obtain a first set of entity embeddings by training the maximum margin model using the Wikidata data for each entity in each of the languages available: label, description, title (we intended to extend this with other Wikidata properties, e.g subclass of, type, etc..). For this we experimented using only the CLS token, the mean pooling of the token embeddings, or the individual embedding of each of the tokens. The better results were obtained by using mean pooling. The negative samples are collected by a random permutation of all the positive embeddings. This step (2) ended up being optional in the training pipeline, during the final reporting period, since results are satisfactory by going directly from step (1) to step (4).
3. Do the same procedure as above using the concatenation of all the Wikipedia language pages for each entity as positive examples. This step (3) also ended up being optional in the training pipeline, during the final reporting period, since results are satisfactory by going directly from step (1) to step (4).
4. Do the same procedure as above using as positives the Wikipedia contexts where there is a link to an entity.
5. A final optional step is to further train the entity embeddings using as positives all the entities that co-occur as a link with another entity.

In the final reporting period, we investigated whether these embeddings were suitable for the downstream task of Entity Linking, by using them in our aforementioned modified DCA model. We trained these embeddings for approximately 20M entities of Wikipedia and were able to train and store them in bf-16 for memory efficiency, without a loss in performance. These embeddings, along with our DCA model, obtain SOTA results on several Entity Linking test sets.

In this final reporting period, we also experimented with adding an additional binary classification head for nil detection, where each selected entity of DCA is classified whether it is or is not a nil mention. For this model, we utilize the local, P(E|m), global and mention type scores of the selected entities along with three other scores: (i) a score based on the difference of the cosine similarity of the selected entity with the remaining candidate entities and the cosine similarity of the selected entity and the remaining set of all entities; (ii) a coherence score of the selected entity with the remaining selected entities in the document; and (iii) a score based on the KL divergence between the distribution of scores of the candidate entities for a mention and the uniform distribution.

In the evaluation section, we report the final results obtained on the quality of the entity embeddings, the entity disambiguation quality of our modified DCA model, and also preliminary nil detection results.

3.3. Story Segmentation

The human voice has a personal identity that may offer biometric security by combining physiological and behavioral characteristics (Lu et al. (2017)). Driven by a great deal of potential applications in story segmentation, automated systems have been developed to automatically extract the different pieces of information conveyed in the speech signal. Hence, several tasks could be defined under the speaker recognition problem. They differ mainly with respect to the decision type that is required for each task. In speaker identification, a voice sample from an unknown speaker is compared with a set of labeled speaker models (Tirumala et al. (2017)). The label of the best matching speaker is taken to be the identified speaker. In a speaker verification task, an identity claim should be provided or asserted along with the voice sample (Nagrani et al. (2020)). The unknown voice sample is compared only with the speaker model whose label corresponds to the identity claim.

A more challenging task is generally referred to as speaker diarization which is used to answer the question of "who spoke when?" (Wang et al. (2018)). Throughout the diarization process, the audio data would be divided and clustered into groups of speech segments with the same speaker identity/label. A complicating factor for this task is that the input news stream may contain speech from more than one speaker. Thus, speaker diarization is regarded as the combination of speaker

segmentation and speaker clustering. The first aims at finding speaker change points in an audio stream and the second aims at grouping together speech segments based on speaker characteristics.

In our initial experiments, we only investigated recognition tasks. Specifically, we focus on text-independent speaker recognition when the identity of the speaker is based on how the speech is spoken, not necessarily on what is being said. Typically, such a system operates on unconstrained speech utterances, which are converted into vectors of fixed length, called speaker embeddings.

Recently, x-vector-based architectures attained state-of-the-art results on speaker-related tasks (Snyder et al. (2018a)). The development of time-delayed neural networks (TDNNs) topology is still an active research area in speech processing. The preferred approach is to train neural networks on the speaker classification task. After the model convergence, low-dimensional embeddings are extracted from the bottleneck layer before the softmax output. Speaker recognition can be completed by comparing the two embeddings over a cosine distance measurement to accept or reject a hypothesis that both samples contain the same speaker. Additional complex backend scoring can also be utilized for this task, such as probabilistic linear discriminant analysis (PLDA) (Ioffe (2006)).

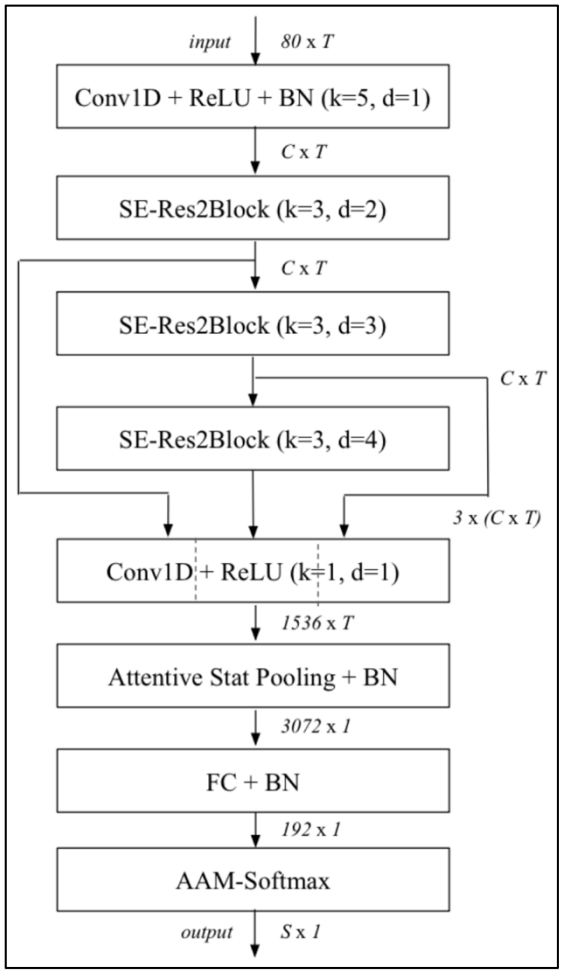

Figure 2 *Network topology of the ECAPA-TDNN (Desplanques et al. 2020) embedding extractor where BN stands for batch normalization and the non-linearities are rectified linear units (ReLU)*

The statistics pooling layer in the x-vector system can map the variable-length input into a fixed-length representation by gathering temporal statistics of hidden layer activations. Okabe et al. (2018) introduced a self-attention system to the statistical pooling, focusing more on essential frames. This model is then improved by adding elements of ResNet architecture (He et al. (2016)). The residual connections of ResNet between the frame-level layers can enhance the x-vector embeddings. Moreover, these residual connections improve the backpropagation in terms of faster convergence and prevent the vanishing gradient problem (Snyder et al. (2018b)).

In this deliverable, we follow ECAPA-TDNN (Desplanques et al. (2020)) architecture which can eliminate some limitations of the x-vector embeddings. This new model extends the temporal

attention mechanism even further to the channel dimension. It enables the network to focus more on speaker characteristics that do not activate on identical or similar time instances. An overview of the complete architecture is given by Figure 2 where *k* and *d* represent kernel size and dilation spacing of the network layers. *C* and *T* correspond to the channel and temporal dimension of the intermediate feature maps, respectively, and *S* is the number of training speakers/users.

Channel- and context-dependent attention mechanisms are implemented inside the pooling layer, which allows the network to attend different frames per channel. The temporal frame context in the original x-vector model is limited to *15* frames (Garcia-Romero et al. (2019)). As the model benefits from a broader temporal context, it is possible to rescale the frame-level features given global properties of the input sample, similar to the global context in the attention modules. Therefore, 1-D squeeze-excitation (SE) blocks (Hu et al. (2018)) rescale the channels of frame-level feature maps to insert global context information inside the locally operating convolutional blocks.

Regular residual blocks (ResBlocks) make it easy to incorporate advancements concerning computer vision architecture (He et al. (2016)). The recent Res2Net module enhances the central convolutional layer such that it can process multi-scale features by constructing hierarchical residual-like connections within (Gao et al. (2019a)). Thus, integrating 1-D SE-Res2Block improves performance while simultaneously reducing the total parameter count by hierarchically used grouped convolutions.

At the last stage, multi-layer feature aggregation (MFA) merges complementary information before the statistics pooling by concatenating the final frame-level feature map with intermediate feature maps of preceding layers (Gao et al. (2019b)). The overall network is trained by optimizing the AAM-soft-max (Deng et al. (2019)) loss on the speaker labels of the training data. The AAM-soft-max is an enhancement compared to the traditional soft-max loss in the context of fine-grained classification problems. It directly optimizes the cosine distance between the speaker embeddings. In this way, complex scoring backends, like PLDA, can be avoided.

3.4. Online News Classification

For the classification of online news, Priberam has worked with the taxonomy established by the International Press Telecommunications Council (IPTC), a consortium of the world's major news agencies. The IPTC Subject Codes vocabulary and the succeeding Media Topics vocabulary establish a hierarchical system of labels to describe the topics covered by any media document. In our experiments, the subject codes vocabulary has been used to classify news articles, and it covers 1404 labels of topics distributed over a hierarchy of three layers. Label names and descriptions are included in seven languages (English, German, French, Portuguese, Spanish, Italian, and Japanese).

Using a dataset of Portuguese news provided by the Lusa News Agency[1], Priberam has trained models for news classification in this taxonomy. The dataset includes over 700,000 news articles in Portuguese for training and testing and an additional 1,000 articles in Spanish and English, each provides a general sense of the cross-lingual performance of the model.

An additional dataset of news articles was acquired from the STT Finnish News Agency[2], the dataset includes over 900,000 news articles in Finnish, labelled with IPTC labels (Politics., Econony, business and finance, etc.) We use this dataset along with the Lusa dataset to train our models in a broader topic space, and to help multilingual models not overfit to a single language. As a Uralic language, Finnish is lexically very distant to Portuguese.

Previous approaches to this task by Priberam used a model described in report D5.1 of the SUMMA project[3], which used convolutional neural networks (CNNs) to aggregate word embeddings to make a final decision through a fully connected layer. Separate versions of this model made decisions at each step of the label hierarchy. For the model to cover languages outside the training set, the FastText (Bojanowski et al. (2017)) multilingual word embeddings were used. The FastText word embeddings were initially published by Facebook research as separate sets of monolingual embeddings for 89

[1] Lusa Agency of Portugal: https://www.lusa.pt/lusanews

[2] STT Finnish News Agency: https://stt.fi/en/

[3] SUMMA Deliverable D5.1: http://summa-project.eu/wp-content/uploads/2017/08/SUMMA_D51_InitialNLU.pdf

languages, these were later aligned by researchers at Babylon Health into a single set of multilingual embeddings. This allows the model to infer on zero-shot languages. These embedding vectors were not fine-tuned in training, which avoids corrupting the word embeddings of languages not seen during training.

One of the main focuses of the news classification task is to improve the performance of Priberam's news classifier. Firstly, by finding a lighter model that can predict the entire label hierarchy in a single forward pass. And secondly, by leveraging the new developments in NLP model architectures, namely models such as bidirectional encoder representations from transformers (BERT) (Devlin et al. (2019)), that can be pre-trained in a multilingual context and then fine-tuned for the specific task using the monolingual dataset.

Chalkidis et al. (2020) performed a thorough survey on the hierarchical multi-label classification of text and showed the outstanding performance of transformer type models. A significant drawback of these models is the limited input size that requires some news articles to be shortened.

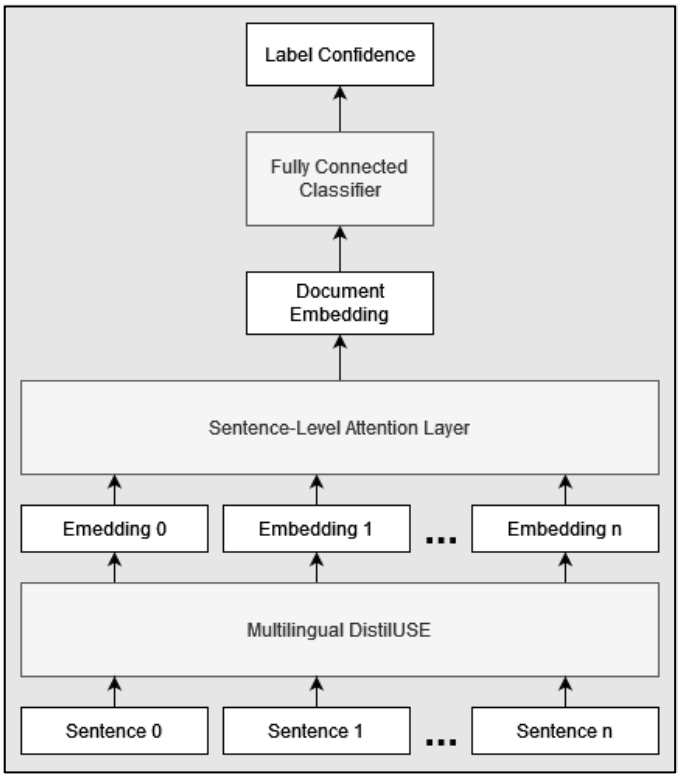

Figure 3 *Architecture of sentence embeddings-based classification models where the novel sentence-level attention layer can take queries from various sources, and outputs an embedding*

Our first proposed model uses multilingual sentence embeddings produced by a DistilUSE (Reimers et al. (2019)) model to represent an entire news article as a sequence of sentence embeddings. DistilUSE is a transformer-type model trained as a more lightweight multilingual counterpart to a monolingual teacher model (using knowledge distillation). This model is trained to generate sentence embeddings in a shared multilingual space. In our new proposed architecture, an attention layer is used to estimate the importance of each sentence embedding and aggregates them for a final decision in a fully connected layer. We further expanded on this practice by experimenting with separate attention queries for each label and particular attention queries for each hierarchy depth. The general architecture of these models is shown in Figure 3.

Our second proposed model is based on an attention-aware model called AttentionXML (You et al. (2019)), which has shown remarkable performance in use-cases of extreme multi-label classification. AttentionXML works by allowing each candidate label to query directly on the word embeddings. The result of this attention layer is fed to a fully connected binary classifier shared between all labels.

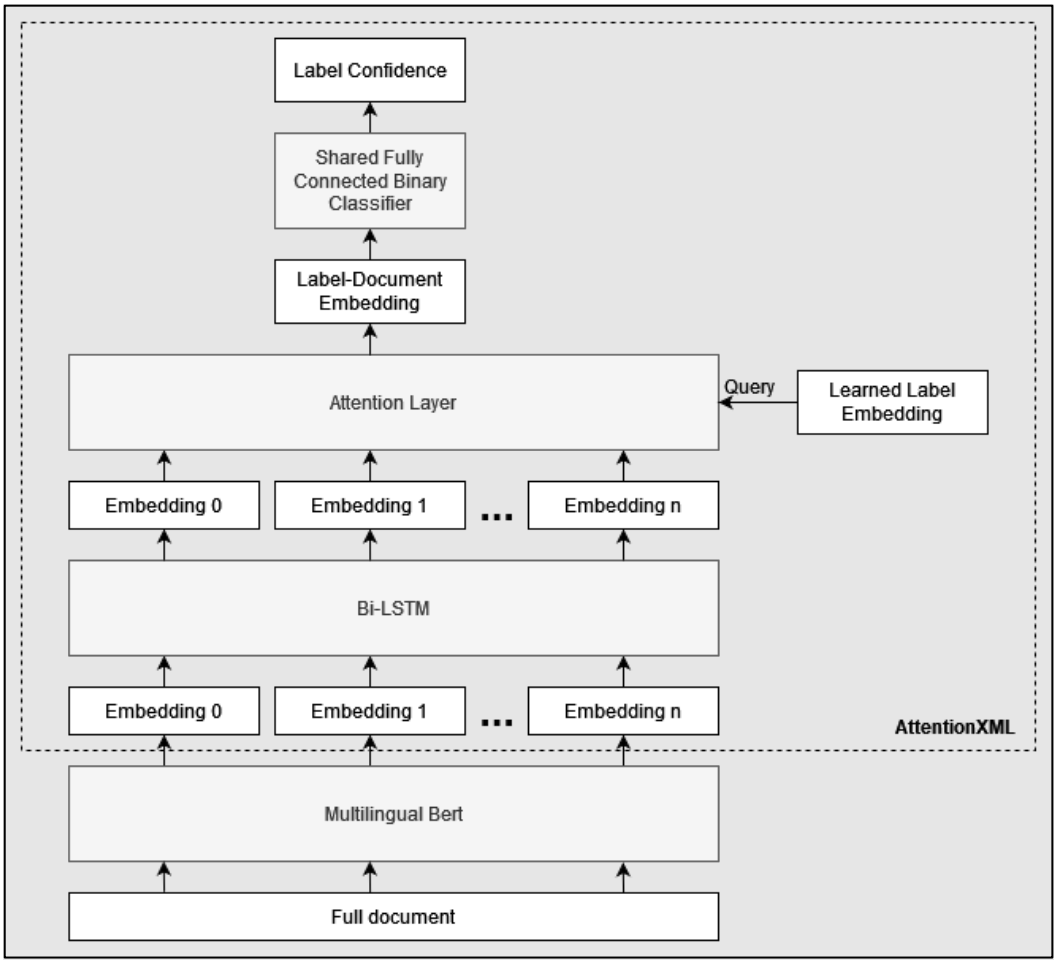

Figure 4 *Architecture overview of mBERT and AttentionXML hybrid models, the top dashed box shows the architecture of a stock AttentionXML*

Each label learns its own query, which finds the most relevant words. The final classification layer is trained on identifying if the document has the most attention on the words that are relevant to its topics. The major drawback of this model is its non-reliance on pre-training and the lack of multilingual support. We explore two modifications to this model, both aimed at making it multilingual. Firstly, we experimented with replacing the word embeddings with pre-trained multilingual word embeddings, and we chose the BPEmb (Heinzerling & Strube (2018)) embeddings for this. These are subword embeddings trained with byte pair encoding that outperform FastText in some scenarios. The authors have open-sourced BPEmb embeddings and tokenizers for 275 languages, along with a multilingual version that covers all 275 languages. Secondly, we try replacing the entire embedding layer with a transformer model. For this, we used a multilingual BERT to

provide contextual embeddings for each token that serves as input to AttentionXML. This allows our embeddings to be more contextualized than what can be achieved with the default biLSTM and will enable us to partially finetune the mBERT model, improving its accuracy for the task without sacrificing the multilingual performance. We later run similar experiments with a pretrained multilingual Roberta-Large model (Liu et al. (2019)), which has shown great potential for multilingual NLP tasks. The architecture of these latter models is shown in Figure 4.

Improving Explainability of AttentionXML-Based Models

When initially proposing the architecture of AttentionXML, the original authors boast about how the model provides a simple explanation for model decisions since it has a single Token to Label attention layer. And that the attention values from this layer provide a score of how relevant each word token is for each label decision.

In our analysis of these attention distributions, we found many examples where the attention peaks were not on the relevant tokens, but instead other tokens in the neighborhood of these relevant tokens. We speculate that the BiLSTM of the model can aggregate information in the contextual embeddings near to the relevant text spans, and that some arbitrary embedding might be sufficient for the model to make a decision. To minimize this effect, we experiment with splitting the BiLSTM, into a separate Forward-LSTM and Backward-LSTM. The reasoning for this is that, since the LSTMs can accumulate relevant information onto arbitrary tokens in the neighborhood of the relevant sections, and that these tokens will later be favored by the attention layer, using separate LSTMs with different directions will restrict the positions at which these tokens will be found. With the Forward-LSTM, we can guarantee that the relevant information can only be accumulated on a token at the end of, or to the right of, the relevant section. Similarly, with the Backward-LSTM, we can guarantee that the relevant information can only be accumulated on a token at the start of, or to the left of, the relevant section. This way it is possible to find relevant spans delimited by the high attention tokens of these two models.

We also experimented with the attention layer of the mBert and AttentionXML hybrid model. Here we found that the attention weights were seemingly very random. Given the size of the transformer-type models, it is perhaps not valid to think of the output embeddings as contextual embeddings of

the corresponding input tokens, in the same way that the output embedding of the [CLS] token is used as an embedding of the entire document.

Experiments with ICD coding

Multi-label classification of medical documents regarding diagnosis and procedures described within medical records is a popular task and benchmark for the models described here, due to its necessity and applicability in hospitals. The International Classification of Diseases (ICD)[4] is a globally used labelling schema, maintained by the World Health Organization (WHO). Due to the very large label space of ICD, and its somewhat hierarchical nature, ICD classification of medical documents is a similar task to IPTC classification. We experiment our models on the widely used MIMIC dataset (Johnson et al. (2016)), which is labelled according to the ICD9 version of the classification standard.

SmartTags: Continuously learning to suggest news articles according to user preferences

We also considered a task where the user defines a category *(tag),* and the system identifies news articles that correspond to the category and continuously learns from user feedback (Mendes, 2023). Specifically, the user creates a tag by providing a textual description of the topic or story of interest and selecting a few relevant documents and the system should suggest more articles relevant to the tag as they arrive. It may also request the user to label specific articles as either relevant or irrelevant to the tag and learn from the user's feedback.

Our approach can be summarized as follows. Given a query document that we want to classify as either relevant or irrelevant for a given tag:

1. Contextual embeddings are obtained for the query document, the tag description, and each of the *n* user-labeled positive documents using the *all-mpnet-base-v2* sentence transformer. More concretely, we obtain an embedding vector for the tag and three embedding vectors for each document: one for the title, one for the first paragraph, and one for the rest of the article (article body).

[4] International classification of diseases (ICD): https://www.who.int/classifications/classification-of-diseases

2. For the query – tag description pair, a 3-dimensional feature vector is obtained where the components correspond to the cosine similarities between the tag description and the query title, the first paragraph and the article body,
3. For each of the *n* query – positive article pairs, a 3-dimensional feature vector is built where the components correspond to the cosine similarities between the titles, the first paragraphs and the article bodies of the two articles.
4. A score s_d is obtained for the query – tag description pair by passing the corresponding feature vector through a linear SVM.
5. *n* scores $s_1, s_2, \ldots, s_n$ are obtained by passing the feature vectors of each query – positive document pair through a linear SVM.
6. The *n+1* scores are ensembled to produce a final score for the query document.
7. A threshold is applied to this score to decide whether the document is classified as relevant or irrelevant to the tag.

The ensemble score in step 6 is obtained using:

$$S = \alpha s_d + (1-\alpha) g(s_1, s_2, \ldots, s_n),$$

where *g* is a permutation-invariant aggregation function and $\alpha \in [0,1]$ is a hyperparameter controlling the relative weight of the two terms. We experimented several possibilities for *g,* such as the sum, the maximum and the Mellowmax (Asadi, 2017) of the scores and also the following probability-based score, which is inspired in the formulation of a conditional random field – see Mendes (2023) for mathematical details:

$$g(s_1, s_2, \ldots, s_n) = \left(1 + \prod_{i=1}^{n} \frac{1-\sigma(s_i)}{\sigma(s_i)}\right)^{-1},$$

being σ the sigmoid function.

Regarding the data, to the best of our knowledge, there is no public dataset that directly fits our problem. To overcome this issue, we used a news dataset provided by DW. The useful feature of this dataset is the fact that each document is accompanied by a list of human-written keywords. We use these keywords to construct synthetic tags and artificially label each document as either positive or negative for each tag. Specifically, we select a set of two keywords and then label as positive all the

documents that contain those keywords in their keyword list and as negative otherwise. Tag descriptions were generated using a Llama2-chat large language model (LLM). The prompt contained a few examples of keywords along with tag descriptions written by humans. This setup enabled few-shot learning for the LLM. We provide some examples of tag descriptions generated by the LLM and the results of our model in section 4.5.

3.5. Online News Clustering

Our primary focus for the news clustering task is to build an online multilingual news clustering system that could process and organize articles from most SELMA languages[5]. In this task, a continuous stream of incoming news articles must be organized into clusters of events called stories. Miranda et al. (2018) approached this problem by processing the news documents stream into monolingual and cross-lingual clusters. Each document is first associated with a monolingual cluster using the term frequency-inverse document frequency (TF-IDF) sub-vectors of words, lemmas, and named entities. Then, cross-lingual clusters are computed by linking different monolingual clusters through cross-lingual word embeddings weighed with TF-IDF. While this approach obtained good results at the monolingual level, it had the following drawbacks: the cross-lingual word embeddings did not take their neighboring words (and thus, the context of the sentence) into account, and the monolingual step required training a separate model for each language as well as extracting the entities from the given text, a task that can be problematic for low-resource languages.

For our approach (Santos et al. (2022)), we developed a system that can cluster news articles of any language without depending on language-specific features while being supported by pre-trained multilingual contextual embeddings. For a given document, our system is composed of four main steps: (i) obtaining its document representations, (ii) finding the best-ranked cluster for that document, (iii) deciding if the document accepts the best-ranked cluster and enters it, and (iv) merging

[5] SELMA platform target languages: Albanian, Arabic, Bulgarian, Chinese, Croatian, English, French, German, Greek, Hindi, Indonesian, Macedonian, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Spanish, Turkish, Ukrainian, Urdu.

clusters that pertain to the same story. A representation of our clustering system is depicted in the following figure.

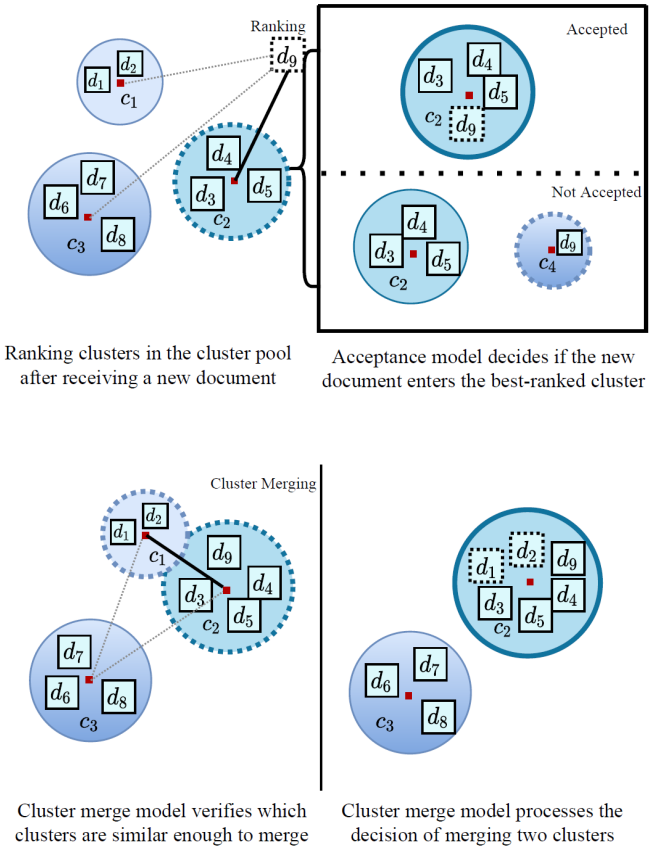

Figure 5 *Representation of the news clustering system's ranking, acceptance and merge steps*

To represent news documents and clusters, we focused our efforts on composing a contextual representation in a set of dense vectors. To that end, similarly to the news classification task, we use DistilUSE, a pre-trained model that aligns text at the sentence level into a shared semantic space, resulting in similar sentences being closely mapped in the vector space regardless of their language. This model supports over 50 languages and does not require the specification of the input language, providing a vectorial representation for the documents that can then be used to inference and group similar news articles. This is a significant change from previous approaches, as contextual information was not taken into account at a cross-lingual level in news clustering state-of-the-art (Miranda et al. (2018), Linger et al. (2020)). Additionally, this approach simplifies the clustering task

by using a single cross-lingual representation for the documents, thus allowing for a fully dense clustering space.

Documents are comprised of two components: a set of dense vectors d^r corresponding to a contextual representation of the document, and a temporal representation (d^{ts}). For each document, d^r contains three dense representations: d^r_1 corresponds to its body and title, d^r_2 to its first paragraph, and d^r_3 to its first paragraph and title. Each of the output vector representations is obtained by mean pooling. Regarding the temporal representation, we follow previous approaches (Miranda et al. (2018)) and expose the temporal representation d^{ts} of a document as the value of its timestamp in days.

In order to find the best-ranked cluster for a given document, we trained a Rank-SVM model, which is a variant of the support vector machine (SVM) algorithm, using a news clustering dataset (Rupnik et al. (2016)) with dense and temporal features. Given the training partition of the dataset, each document generates a positive example corresponding to its gold cluster, and 20 negative examples for the 20 best-ranked clusters that are not the gold cluster.

These examples are then used in the Rank-SVM to obtain a set of fixed weights for each feature. Temporal features are computed through the Gaussian similarity between two timestamps (represented by the $score^{ts}$ function, and the dense features are obtained through the computation of the cosine similarity ($score^{cos}$). The ranking score of a cluster *c* given a document *d* and the ranking model's fixed weights *u* is formalized as follows:

$$score^{rank}(d,c) = \sum_{i=1}^{3} \left(score^{cos}(d_i^r, c_i^r) \cdot u_i^r\right) + \sum_{j=1}^{2} \left(score^{cos}(d_{j+1}^r, c_1^r) \cdot u_{j+3}^r\right) + \sum_{k=1}^{3} \left(score^{ts}(d^{ts}, c_k^{ts}) \cdot u_k^{ts}\right)$$

After computing the best-ranked cluster *c* for a given document *d*, a trained SVM model, which we refer to as the acceptance model, determines if the document enters the cluster by computing its acceptance score, represented as follows (*v* corresponds to the acceptance model's weights):

$$score^{accept}(d,c) = \sum_{i=1}^{3}\left(score^{cos}(d_i^r, c_i^r)\cdot v_i^r\right) + \sum_{j=1}^{2}\left(score^{cos}(d_{j+1}^r, c_1^r)\cdot v_{j+3}^r\right) + \sum_{k=1}^{3}\left(score^{ts}(d^{ts}, c_k^{ts})\cdot v_k^{ts}\right) + score^{rank}(d,c)\cdot v^{rank}$$

Finally, after receiving a new document, a cluster verifies its similarity with each cluster in the cluster pool using the ranking model described above. Each candidate cluster is then evaluated by a third SVM model, which we call cluster merge model, and the documents from each cluster evaluated as a positive match are inserted into the source cluster. The intuition for this model is to find separate clusters that pertain to the same story and subsequently merge them. This may happen throughout the clustering process; since few documents related to a given story have entered the system, the acceptance model may mistakenly assign separate clusters to those documents initially. As more relevant documents enter the system, those clusters may end up in similar points in the vector space and thus should be merged.

3.6. News Summarization

Monolingual Abstractive Summarization

Text summarization aims at producing a short text segment that preserves the essential information conveyed by a longer source document. The approaches for automatic summarization can be divided into two categories: extractive and abstractive methods. The former address the problem by identifying salient parts of the source document and directly copying those to the summary (e.g., Kupiec et al., 1995, Dorr et al., 2003, Nallapati et al., 2017). The latter produce the summary by generating new text that paraphrases the most relevant parts of the source document (e.g., See et al., 2017, Guo et al., 2018, Lewis et al., 2020).

In SELMA, our research focused on summarizing video transcriptions using current neural approaches. Since extractive methods produce weak summaries over automatic transcriptions (given the low quality of the generated sentence boundaries), we shift toward abstractive summarization methods. Nonetheless, abstractive summaries often contain factual inconsistencies that hamper the

adoption of these approaches in practical applications (Kryściński et al., 2019a). For this reason, our main goal is to develop techniques to enhance the factual consistency of the generated summaries.

Our work (Pernes et al., 2022) builds upon the state-of-the-art methodologies for abstractive summarization, namely those based on transformer sequence-to-sequence architectures, like BART (Lewis et al., 2020), a pre-trained encoder-decoder transformer that can be finetuned in a wide range of text generation tasks, including summarization. At the same time, automatic evaluation metrics such as CTC scores (Deng et al., 2021) have been recently proposed that exhibit a higher correlation with human judgments than traditional lexical-overlap metrics such as ROUGE. In our work, we close the loop by leveraging the recent advances in summarization metrics to create quality-aware abstractive summarizers. Namely, we proposed an energy-based model that learns to re-rank summaries according to one or a combination of these metrics. An overview of the proposed framework is presented in Figure 6. As suggested by the picture, the energy-based re-ranking model (EBR) is presented with a set of candidate summaries for a given source document and assigns a score to each candidate. The EBR is trained to mimic the ranking induced by a pre-specified gold-metric, so that the scores it provides should indicate which candidate is the best one according to that metric. We experiment using several metrics to train our energy-based re-ranker and show that it consistently improves the scores achieved by the predicted summaries. Nonetheless, human evaluation results show that the re-ranking approach should be used with care for highly abstractive summaries, as the available metrics are sometimes not sufficiently reliable for this purpose.

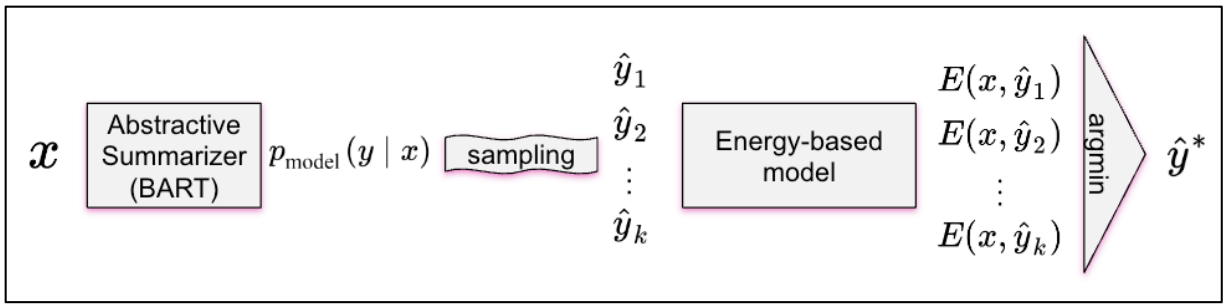

Figure 6 *Representation of the proposed energy-based re-ranking approach for abstractive summarization.*

Cross-Lingual Abstractive Summarization

Our previous work has focused only on abstractive summarization of English documents. Multilingual summarization resources are scarce, not only in terms of availability of trained models, but also in terms of public datasets and evaluation metrics. Furthermore, SELMA proposes to deal with multilingual data streams, and therefore it would not be appropriate to follow monolingual approaches to summarization. Therefore, we address the challenging problem of cross-lingual summarization, where a summary in each target language is generated from a source document in another language. Moreover, if information is not conveyed consistently across languages, the trustworthiness of the system is compromised. Users cannot rely on the summaries to be accurate and unbiased, regardless of the language in which they consume the content. Therefore, in the context of SELMA, we have proposed and addressed the task of multi-target cross-lingual summarization, which imposes the additional constraint that summaries in different languages for the same document should convey identical information.

We have not been able to publish or preprint our work at the time of writing, but our main contributions to the topic follow. Apart from the task itself, which is novel and reflects an aspect that should be considered by reliable multi-target cross-lingual summarization systems, we propose a set of strong methods for this task and establish appropriate evaluation metrics that measure semantic consistency across multiple target languages.

An obvious solution to this problem is to take a language as a pivot and use a pipeline of summarization and translation models. Specifically, we first produce a summary in the pivot language and then translate it to each of the desired target languages. Since translation is a more deterministic task than summarization, the resulting summaries should be semantically very similar to each other. However, it has inherent drawbacks. It involves two non-parallelizable phases of decoding: first generating the pivot summary, and then generating summaries for each target language. This approach can also suffer from error accumulation from both decoding phases. Moreover, it exhibits a bias towards the pivot language, which may reflect biases introduced during the summarization to the pivot language in the resulting target languages.

Another option proposed in our work involves choosing candidate summaries based on their semantic similarity. For this purpose, we used a state-of-the-art text encoder to produce cross-lingual

embeddings for each summary and then re-ranked candidates to obtain a set of summaries with high cross-lingual similarity. We propose both a pivot-dependent and a pivot-free approach. The latter uses dynamic programming to find a set of summaries in different languages with high semantic similarity. Notably, this approach mitigates the bias introduced by the choice of pivot language and allows for higher GPU parallelization as all summaries can be decoded in parallel. Figure 7 represents each of the three approaches as graphical models. We use x_o to denote the source document, y_π to denote the pivot summary written in the pivot language π, y_t to denote the summary in each target language t, and N to denote the number of target languages.

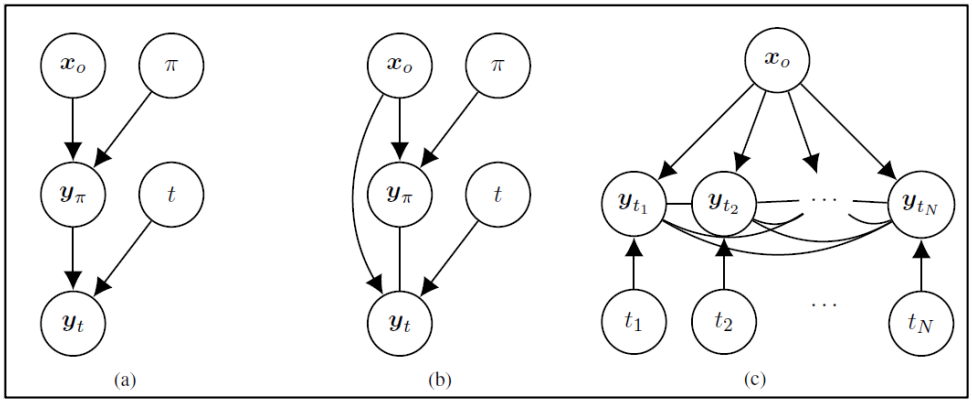

Figure 7 *Graphical models representing summarize-and-translate (a), our pivot-dependent (b) and pivot-free (c) approaches*

As for the evaluation metrics, apart from computing ROUGE scores with respect to the reference summaries in each target language, which to some extent allow to measure the relevance of the generated text, we also consider reference-free machine translation metrics, that can be used to measure the degree of semantic consistency across target languages.

Multilingual Multi-Document Extractive Summarization

The centroid method is a simple approach for extractive multi-document summarization and many improvements to its pipeline have been proposed. Our work (Gonçalves et al., 2023), builds upon the contribution of Ghalandari et al. (2019) by adding a beam search process to the sentence selection and a centroid estimation attention model that leads to improved results. Notably, by using cross-lingual sentence embedding models to generate cluster representations, our approach is applicable in

a truly multilingual setting, where a cluster of documents to be summarized may contain documents in different languages.

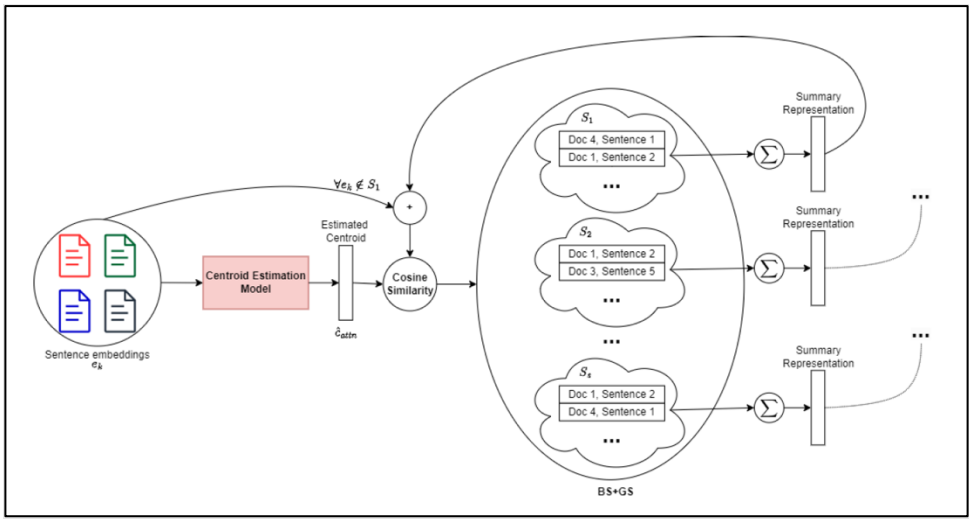

Figure 8 *Outline of the proposed methodology for multilingual, multi-document extractive summarization*

The approach is outlined in Figure 8. First, cross-lingual sentence embeddings are computed for each sentence in the source document. Then, instead of representing the cluster by the average of the sentences within it, the representation we use as the centroid is obtained by feeding the set of sentence embeddings into an attention-based model that is trained to approximate the embedding obtained by averaging the sentences of the cluster reference summary. In this way, the model implicitly learns to give higher weight to the most relevant sentences within the cluster. After this centroid estimation step, a beam search is used in the sentence selection step to select the sentences with maximum similarity to the estimated centroid, which form the extractive summary.

This model is evaluated on a set of standard benchmark datasets and outperforms the original unsupervised approaches in most cases. Due to the lack of multilingual, multi-document summarization datasets, we adapted the CrossSum dataset (Bhattacharjee et al., 2023), originally designed for cross-lingual abstractive summarization, to our task, which is another contribution of our work.

Speech Summarization

This task is conceptually identical to text summarization except for the input modality, which is now raw speech instead of text. In the context of SELMA, speech summarization plays a central role, since our efforts in abstractive summarization have the ultimate purpose of summarizing videos.

This task is traditionally divided into two independent subtasks: automatic speech recognition (ASR), which produces the audio transcripts, and text summarization, which produces the summary given the transcript. However, the ASR step leads to error propagation and loss of the information provided by the speaker's intonation. As a first step to address these problems, we developed an abstractive summarization system capable of performing end-to-end speech-to-text summarization, i.e., without an intermediate transcription step.

Our approach (Monteiro and Pernes, 2022) uses a pre-trained wav2vec 2.0 model (Baevski et al. 2020) to extract audio embeddings from the raw waveform and a transformer decoder to generate the summary text. This decoder is taken from a transformer previously trained on the text-to-text summarization task, so the decoder expects to receive textual token representations rather than the audio embeddings provided by wav2vec. Therefore, we trained an LSTM-based cross-modality adapter on the task of converting sequences of audio embeddings into the corresponding sequences of textual embeddings. Specifically, given an audio and its corresponding transcript, we extract the embeddings from the audio using the word2vec model and the corresponding textual embeddings using the encoder of the transformer trained for textual summarization. We then use these pairs of audio and text embedding sequences to train the cross-modality adaptor. By cascading the wav2vec encoder, the cross-modality adapter, and the transformer decoder, we obtain an end-to-end model for speech summarization. Figure 9 illustrates an overview of both our end-to-end approach and the cascade methodology, detailing their respective architectures.

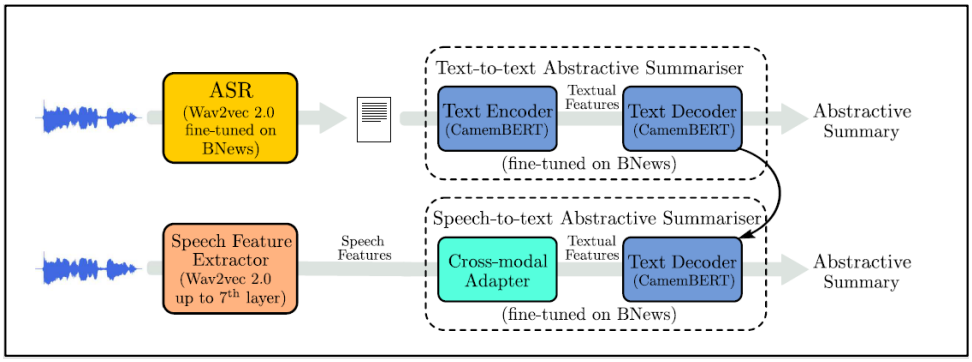

Figure 9 *Architectures of the cascade and end-to-end systems for speech-to-text summarization*

We remark that surpassing the performance of the cascaded (ASR + text summarization) approach is difficult given the relatively limited amount of data available. To narrow this gap, we conducted a pre-training of the cross-modal adapter using ASR data in three steps, as described next. The input speech features and target textual features were normalized such that each dimension had zero mean and unit variance.

Stage 1: At this stage, we used the same Common Voice corpus that was used to train the ASR model. A proportion of speech features from the input sequence is randomly masked. The cross-modal adapter is trained to minimize the mean squared error (MSE) between the reference embeddings and the predicted embeddings.

Stage 2: We dropped the Common Voice dataset and used our corpus during this training stage. The objective remains to minimize the MSE. Masking is no longer used and the default teacher forcing algorithm for training seq2seq models is replaced by a peeling back strategy.

Stage 3: The cross-modal adapter is now trained to predict the end of the sequence of textual embeddings, again using our dataset. Given a sequence of predicted textual embeddings, predicting time step whether it is the end of the sequence is a binary classification problem. Thus, minimizing binary cross-entropy loss suffices.

After this three-stage pre-training, the cross-modal adapter and the text decoder are jointly trained for abstractive summarization using our dataset in a multitask objective consisting of the usual cross-entropy loss for summarization and the binary cross-entropy for EOS detection.

4. Experimental Results

This section includes the experimental analysis of the previously defined problems, alongside their discussions. Sub-section titles are arranged in accordance with the previous section.

4.1. Named Entity Recognition

Hierarchical Nested NER

We report named entity recognition and classification (NERC) F1 scores obtained for all entities. For each level of the hierarchy, we utilize two internal datasets related to media content: i) MediaPT, containing *42,000* training examples in Portuguese; and ii) MediaDE, containing *85,000* training examples in German. Both datasets have the same set of 61 labels, including hierarchy levels, e.g., "gpe → administrative_region → municipality", where "gpe" is the top-level (L0), "administrative_region" corresponds to L1, and "municipality" to L2. The obtained results can be seen in Table 1. It is possible to observe that both models achieve similar scores for both languages, with a slight advantage of the stack-LSTM model in MediaPT and the biaffine model in MediaDE. When comparing these models in terms of computational performance, the biaffine approach offers a clear advantage when decoding on CPU or when the sentences are short, with stack-LSTM performing similarly on GPU and slightly better for longer sentences.

Approaches \ Datasets	**MediaPT**	**MediaDE**
Development Set - NERC F1 - ALL (L0 / L1 / L2)		
Stack-LSTM	85.8 (86.5 / 85.4 / 64.5)	80.8 (80.5 / 82.4 / 59.2)
Biaffine	85.6 (86.3 / 85.2 / 64.4)	81.0 (80.2 / 83.4 / 58.9)
Test Set - NERC F1 - ALL (L0 / L1 / L2)		
Stack-LSTM	85.2 (86.0 / 84.6 / 42.0)	81.7 (81.7 / 82.8 / 53.4)
Biaffine	84.7 (85.7 / 84.0 / 48.4)	81.8 (81.8 / 82.7 / 55.6)

Table 1 *Stack-LSTM and biaffine results for MediaPT and MediaDE development and test sets*

The results presented in Table 1 highlight the previously mentioned advantage of working with pretrained multilingual contextual embedding models, which allows us to train models for different languages, as we did for MediaDE and MediaPT, and to train a single model for several languages. This allowed us to participate in the SlavNER shared task, part of the 8th Balto-Slavic NLP, where our biaffine approach was able to outperform all the other submissions for the NER subtask (Ferreira et al. (2021), Piskorski et al. (2021)), which included nested non-hierarchical entities for six different languages.

Cross-Lingual Hierarchical Nested NER

In the second reporting period, datasets for additional languages became available, allowing us to further investigate the multilingual capacities of our models. Currently, the project has created NER datasets for English, French, German, Latvian, Spanish, and Portuguese. Additionally, smaller datasets to evaluate language transfer are also available for Ukrainian, Dutch, and Turkish. Table 2 contains the description of the datasets annotated with our ontology as described in D6.1 - Initial Data Management Plan. The number of annotations shown in the table are counted one for each level of the ontology, so the number of different annotated text spans is much smaller than the number presented. English, French, German, and Spanish have a very good coverage, Latvian proved to have sufficient training data, and Dutch, Ukrainian, and Turkish are used only for evaluation proposes.

Language	#documents	#tokens	#annotations*
English	4500	5584365	3103202
French	3003	3086488	1771902
German	3122	2934042	1625036
Latvian	741	573731	321594
Spanish	2576	2855692	1556536
Portuguese	3199	2317747	1283964
Dutch	50	41193	22966
Ukrainian	211	160163	90865
Turkish	100	70815	40967

Table 2 *Multilingual NER datasets, *number of annotations counting with the hierarchy*

In order to compare the multilingual model performance against the monolingual baseline, we retrained all models using xml-roberta-base, except for French where we kept using camembert-base. This option permits verifying whether the multilingual model would outperform a good pure monolingual model. All the models were trained using the stack-LSTM approach with the hyper-parameters selected in our initial experiments. The monolingual results are presented in Table 3, we report F1 values for each of the ontology levels and a global F1 over the complete hierarchy. The global F1 includes the detection of the modifier tags (e.g. *nominal*, *function* and *relation),* which makes this dataset much harder than other datasets publicly available.

Language	**F1**			
	All	L1	L2	L3
English	81.0	82.1	79.7	64.4
French*	85.7	87.2	83.9	0.77
German	82.2	81.9	84.0	71.5
Latvian	84.2	85.9	82.4	51.1
Spanish	83.5	85.4	81.0	54.5
Portuguese	84.4	85.5	83.4	50.4

Table 3 *Results on test sets training monolingual. *was trained using camembert instead xlm-roberta-base*

For training the multilingual model, we selected English, French, German, Latvian, Portuguese, and Spanish, because they have a good amount of training data. Table 4 reports the F1 values obtained when training the stack-LSTM model with the same hyper-parameters as in the monolingual setting. In this experiment we did not use any artifact to distinguish the languages when training or testing, because this is the simplest, less costly in resources and the most language-independent of the approaches that we researched. Table 4 shows that by training with all languages together we achieve robust improvements in most of the languages except for French, where in the monolingual setting,

we used a base monolingual model (camembert-base). Although the drop of 0.6 is significant, it is not enough to justify the overhead of using a different model in a production scenario.

Language	F1				
	All	Diff to monolingual	L1	L2	L3
English	81.7	**+0.7**	82.5	80.8	64.1
French	85.1	-0.6	86,4	83.4	80.9
German	82.2	**+0.0**	81.7	83.4	72.3
Latvian	85.2	**+1.0**	86.0	84.6	75.0
Spanish	84.4	**+0.9**	86.1	82.3	59.3
Portuguese	85.1	**+0.7**	85.9	84.3	51.8

Table 4 *Results on test sets training multilingual*

We evaluated our model zero-shot capabilities on languages present in the base model but for which we did not have NER training data. Surprisingly and against our best expectations, the multilingual model performs very well on unseen languages. To evaluate the zero-shot setting we asked the annotators to correct, remove, and add to the annotations proposed by the multilingual model. We are aware that this procedure will impose a bias on the annotators leading them to probably keep the annotations of the model, but the cost and feasibility of the task imposes a pragmatic approach. Using those corrected datasets, we evaluated F1 results of the model when seeing the corrected data. Table 5 shows F1 results on the evaluation datasets for Dutch, Ukrainian, and Turkish showing that the annotators did not change much of the annotations proposed by the model for Dutch and Ukrainian. Turkish results have a considerable drop when comparing with the other two languages. This can be justified either by: the quality of the base model (xlm-roberta-base) for Turkish; a real difference in the language itself; or a different criterion was used by the Turkish annotator. If we arrive to the conclusion that the annotation is sound, then we will extend the Turkish dataset and include it in the training data. To further validate these results, we asked Priberam linguists' team to validate each of

these datasets with the annotators, making sure that the applied criteria were the same between these annotators and the original guidelines used for the other languages.

Language	F1			
	All	L1	L2	L3
Dutch	91.4	89.8	93.5	100
Ukrainian	90.8	88.1	94.5	100
Turkish	74.6	71.9	79.9	33.3

Table 5 *Zero-shot results after correcting the annotations predicted by the multilingual model by human annotators*

Lastly, on Table 6 we present the aggregated F1 values with their support on the test dataset for each class on the ontology.

Class	Support	Precision	Recall	F1
animal	31	0.5588	0.6129	0.5846
currencies	409	0.95	0.9756	0.9626
disciplines	127	0.5938	0.5984	0.5961
event	1599	0.7642	0.7073	0.7347
event->festivity	87	0.8144	0.908	0.8587
event->happening	50	0.8049	0.66	0.7253
event->organized_event	869	0.7908	0.7089	0.7476
facility	684	0.7147	0.6667	0.6899
gpe	8659	0.882	0.921	0.9011
gpe->address	86	0.65	0.7558	0.6989
gpe->administrative_region	1494	0.7221	0.7289	0.7255
gpe->administrative_region->municipality	65	0.4921	0.4769	0.4844
gpe->administrative_region->parish	48	0.697	0.4792	0.5679
gpe->city	2257	0.7613	0.852	0.8041

gpe->continent	298	0.7975	0.8591	0.8271
gpe->country	4199	0.9208	0.9586	0.9393
gpe->non_administrative_region	498	0.5365	0.4137	0.4671
gpe->union_of_countries	100	0.899	0.89	0.8945
human_group	536	0.6115	0.4552	0.5219
human_group->ethnicity	67	0.6735	0.4925	0.569
human_group->religion	77	0.7083	0.8831	0.7861
human_work	1116	0.7137	0.6478	0.6792
internet_address	193	0.8585	0.9119	0.8844
internet_address->email	14	0.875	1	0.9333
internet_address->url	80	0.75	0.825	0.7857
location	417	0.7419	0.7098	0.7255
location->astronomical_object	83	0.9067	0.8193	0.8608
location->geographical_feature	99	0.6893	0.7172	0.703
location->river	63	0.78	0.619	0.6903
location->sea/ocean	34	0.8286	0.8529	0.8406
mod-collective	496	0.5718	0.4819	0.523
mod-function	1190	0.7974	0.8832	0.8381
mod-negation	6	0	0	0
mod-nominal	3954	0.6746	0.6396	0.6566
mod-relation	2449	0.866	0.9212	0.8928
mod-sentiment_negative	10	0.8	0.4	0.5333
mod-sentiment_positive	6	0	0	0
number	45	0.7381	0.6889	0.7126
number->license_plate	5	0	0	0
number->telephone	40	0.7381	0.775	0.7561
organization	9431	0.8397	0.8657	0.8525
organization->commercial_company	2323	0.7506	0.7891	0.7694
organization->commercial_company->brand	890	0.6762	0.664	0.6701
organization->cultural_institution	32	0.4865	0.5625	0.5217
organization->education_institution	139	0.7325	0.8273	0.777
organization->educational_institution	160	0.6966	0.6312	0.6623

organization->governmental_institution	2348	0.8176	0.8113	0.8145
organization->healthcare_institution	89	0.71	0.7978	0.7513
organization->intergovernmental_organization	98	0.9255	0.8878	0.9062
organization->media	1562	0.844	0.8867	0.8648
organization->non_governmental_organization	912	0.666	0.7632	0.7113
organization->political_organization	396	0.862	0.8359	0.8487
organization->religious_organization	4	0.5	0.5	0.5
organization->sports_organization	1701	0.8418	0.8601	0.8508
other	725	0.6624	0.5683	0.6117
pathology	671	0.8967	0.9314	0.9137
pathology->disease	434	0.9057	0.9516	0.9281
pathology->pathogen	242	0.8659	0.8802	0.873
people	11937	0.8889	0.9109	0.8998
people->alias	149	0.7857	0.443	0.5665
people->job	3211	0.7494	0.7991	0.7735
quantity	2555	0.9128	0.9346	0.9236
quantity->age	652	0.9109	0.9095	0.9102
quantity->currency	660	0.9207	0.9682	0.9439
quantity->measure	349	0.8015	0.8911	0.844
quantity->percentage	872	0.9479	0.9599	0.9538
quantity->temperature	47	0.8125	0.8298	0.8211
temporal_expression	4725	0.7821	0.8167	0.799
temporal_expression->date	1909	0.9177	0.9466	0.9319
temporal_expression->datehour	185	0.887	0.8486	0.8674
temporal_expression->frequency	142	0.6757	0.7042	0.6897
temporal_expression->hour	372	0.9096	0.9462	0.9275
temporal_expression->period	1116	0.6403	0.6747	0.6571
temporal_expression->time	1897	0.711	0.7612	0.7352
time	1689	0.7824	0.7981	0.7902
time->date	680	0.9137	0.9338	0.9236
time->datehour	2	0	0	0

time->frequency	43	0.6364	0.4884	0.5526
time->hour	154	0.8805	0.9091	0.8946
time->period	323	0.5851	0.7028	0.6385

Table 6 *Full NER results for all languages for each ontology level*

Example-Based NER

The results of example-based NER can be seen in Table 7, where we show the performance for both single-*k* and multi-*k*, for 7 datasets, including different domains, number of training examples, and number of labels. We perform hyperparameter tuning for each dataset using its development set. Few-NERD is the dataset that is more positively impacted by this approach. We hypothesize this could be due to the fact that this dataset is the only one that uses an IO-encoding, which could make it simpler to retrieve the correct tag, as it has to match only the I tag and not the B/I-tags.

Approach \ Dataset	**Few-NERD**	**Onto Notes**	**Co NLL**	**WNUT**	**MIT-R**	**MIT-M**	**ATIS**	**Avg**
Domain	*Generic*	*Generic*	*News*	*Soc. Media*	*Reviews*	*Reviews*	*Dialogue*	-
Trn. Examples	131,000	60,000	14,000	3,400	6,900*	6,700*	6,500	-
# of Labels	66	18	4	6	8	12	68	-
Development Set - NERC F1								
Class. Model	68.31	88.26	95.86	64.75	81.96	73.43	98.19	81.54
+ single-*k*	68.64	88.5	95.86	64.62	82.02	73.61	98.39	81.66
+ multi-*k*	68.75	88.53	95.87	64.74	81.9	73.6	98.33	81.67
Test Set - NERC F1								
Class. Model	67.83	90.11	92.28	57.53	80.05	71.22	95.88	79.27
+ single-*k*	68.18	90.04	92.35	57.61	80.06	71.26	95.86	79.34
+ multi-*k*	68.23	90.08	92.35	57.41	80.22	71.31	95.86	79.35

Table 7 *Example-based NER approach results with single k and multi k for different datasets (*original training data was split into training/validation splits)*

There are cases where development set improvements do not result in test set improvements (OntoNotes and ATIS), or where the improvements in the test set are rather small (remaining datasets). Regarding the possibility of using this approach as a way of incorporating user feedback, we report an experiment where we plot the performance of the linear classifier, the performance of the linear classifier plus kNN using all the available data, and the previous best linear classifier at a certain point plus kNN using the available data (i.e., at point 0.8 we interpolate the predictions made by a linear classifier trained on 60% of the training data, leveraging 80% of the training data as support data). As we can observe in Figure 10 the more support data available for the Few-NERD dataset, the clearer the benefits of using the kNN approach. In particular, it is also possible to observe the slight benefit from continuously collecting data (e.g., by comparing the point 0.8 of the line "Linear Classifier" and the point 1.0 of the line "Previous Linear Classifier + kNN", which only differ in the amount of available support data).

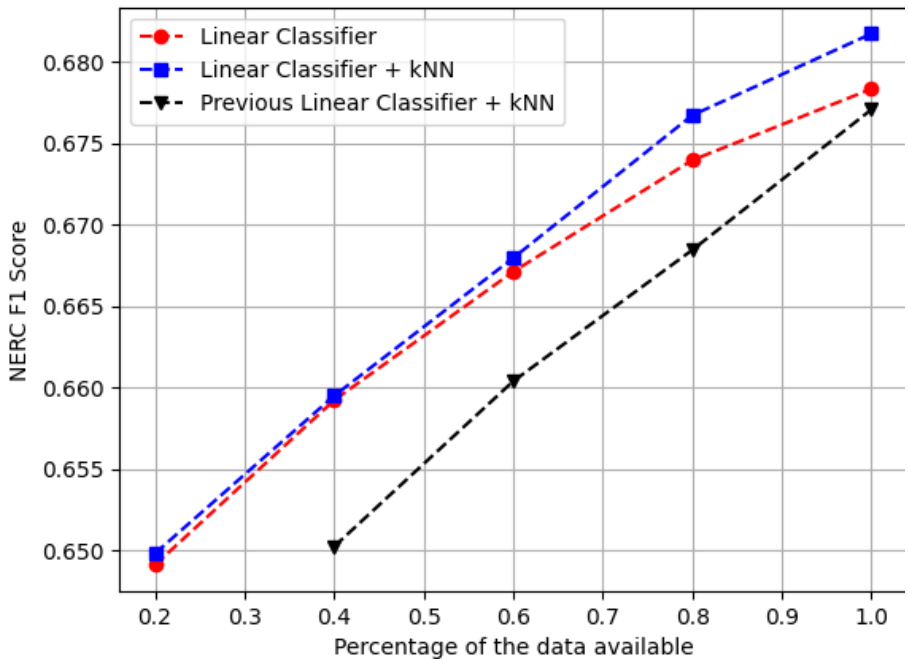

Figure 10 *Impact of increasing support data on example-based NER for the FewNERD dataset*

4.2. Entity Linking and Cross-Lingual Stream Representations

We compare Yang et al. (2019) DCA model with our extended version using multilingual embeddings. We report the in-knowledge-based accuracy (i.e., accuracy disregarding predictions that do not exist in the knowledge base) for several datasets: (i) the English CoNLL 2003 shared task data, containing one development set (Aida-A) and a test set (Aida-B) with news stories from Reuters; (ii) WNED, a collection of English datasets containing news reports and newswire from various agencies (MSNBC, ACE2004, and AQUAINT) or varied English texts such as web pages or Wikipedia pages (CLUEWEB, WIKIPEDIA); (iii) sVoXel (Rosales-Méndez et al. (2018)), a collection of 15 manually annotated news articles, each available in 5 different languages.

Table 8 shows improvements across CoNLL for our base English-only model, but performance on the WNED datasets does not always improve, where the model achieves lower scores, particularly in the CLUEWEB and WIKIPEDIA datasets that are not news related.

Model	Aida-A	Aida-B	MSNBC	AQUAINT	ACE 2004	CLUEWEB	WIKIPEDIA
Original DCA	0.9003	0.8988	0.9334	0.8601	0.8773	0.7634	0.7623
Ours: EN - CoNLL	0.9195	0.9114	0.9395	0.8363	0.8853	0.7564	0.7383
Ours: All - CoNLL	0.9141	0.9157	0.9273	0.8000	0.8773	0.7206	0.7164
Ours: All - Wiki	0.8266	0.8606	0.9288	0.8965	0.8933	0.7515	0.7457

Ours: All - Both	0.8982	0.8921	0.9396	0.8769	0.8853	0.7539	0.7605
Ours: Multilingual Contextual, 300k entities	0.9346	0.8635	-	-	0.8720	0.6936	-
Ours: Multilingual Contextual, 20M entities	0.9380	0.8704	-	-	0.8680	0.6981	-

Table 8 *In-KB accuracy for English datasets for original DCA model and our embedding vocabulary - train data configurations*

Increasing the entity vocabulary leads to a small drop in performance in the WNED collection datasets. Finally, training on Wikipedia leads to a drop in CoNLL performance that can be countered by mixing both train datasets to obtain performance similar to the model using English entities only. This seems to indicate that having training data from different domains (news and Wikipedia) helps the model be more resistant to domain changes. We also report our final multilingual contextual model, trained for 20M entities of Wikipedia across 39 languages. The dataset to train our DCA in this setting was aida-train (CoNLL). We observe a significant increase in results in the Aida-A dataset, while results in Aida-B, ACE, and CLUEWEB did not see an improvement. However, in the multilingual setting of Table 9, the results of this model see a big improvement.

Table 9 and 10 show results for the multilingual scenario in sVoxEL test datasets and the test sets of TAC 2016. The results show that our multilingual model with improved DCA and 20M entities across 39 languages consistently surpasses previous models by a significant margin.

Model	sVoxEL-fr	sVoxEL-de	sVoxEL-it	sVoxEL-es	sVoxEL-en
Original DCA	0.9200	0.8434	0.9173	0.8750	0.9327
Ours: EN - CoNLL	0.9500	0.8737	0.9523	0.9100	0.9625
Ours: All - CoNLL	0.9300	0.8737	0.9474	0.9050	0.9277
Ours: All - Wiki	0.9300	0.8789	0.9373	0.9100	0.9476
Ours: All - Both	0.9300	0.8789	0.9474	0.9100	0.9526
Ours: Multilingual Contextual, 300k entities	0.9402	0.8756	0.9402	0.9303	0.9651
Ours: Multilingual Contextual, 20M entities	0.9402	0.8756	0.9552	0.9303	0.9651

Table 9 *In-KB accuracy in a multilingual scenario for original DCA model and our embedding vocabulary - train data configurations*

TAC Split	Ours: Multilingual Contextual, 300k entities	Ours: Multilingual Contextual, 20M entities
En-News	0.9000	0.9082
En-Discussion Forums	0.8851	0.8752
Es-News	0.9180	0.9279
Es-Discussion Forums	0.8986	0.8920
Zh-News	0.8409	0.8479
Zh-Discussion Forums	0.8176	0.8802

Table 10 *In-KB accuracy in a multilingual scenario for original DCA model and our embedding vocabulary - train data configurations*

Contextual Entity Representations

In this subsection, we report results on the quality of the contextual entity representations. With this aim, we follow the same procedure as Ganea and Hoffman 2017 by computing entity relatedness scores on the dataset from Ceccarelli et al. 2013. We use the same evaluation metrics: normalized discounted cumulative gain (NDCG) and mean average precision (MAP). Table 11 shows our results both for the initial multilingual scenario with mbpe embeddings and using the contextual embeddings of the fine-tuned xml-roberta-base on the multilingual SELMA NER dataset. This also includes the results for our final entity embeddings trained on 20M entities across 39 languages, trained on bf16, which got improved results in all metrics. In Table 12, we show recall results for these embeddings, which were computed by ordering the candidates of mentions in Aida-B by similarity between the corresponding entity embedding and the mean pool of the xml-roberta-base representation of the mention with some context window. We can observe that we obtain a very high recall with this simple method, when we take the top-30 entities for each mention, which indicates that these representations are well suited for the disambiguation task.

--	NDCG@1	NDCG@5	NDCG@10	MAP
Yamada (2016) **English only**	0.59	0.56	0.59	0.52
Ganea and Hoffman **English only**	0.632	0.609	0.641	0.578
Ours multilingual mbpe	0.641	0.604	0.635	0.572
Ours multilingual contextual, mean pool	0.649	0.603	0.629	0.569
Ours multilingual contextual, mean pool (bf16, 20M)	**0.6696**	**0.6238**	**0.6499**	**0.5868**

Table 11 *Entity relatedness on the test set of Ceccarelli et al. 2013*

	Recall@1	Recall@3	Recall@5	Recall@10	Recall@30
Ours multilingual contextual, mean pool (bf16, 20M)	0.4027	0.5740	0.6494	0.7559	9751

Table 12 *Recall@K results on Aida-B of a simple similarity comparison between the entity embedding and the mention*

Nil Detection

In this subsection, we show in Table 13 the nil detection results we have obtained with the approach described in 3.2, on TAC 2016 test sets. This model was trained not only in aida-train of the CoNLL dataset but also on the datasets of TAC 2015. We observe that the F1 results we obtain are particularly satisfactory in the test sets that use the news domain. We further observe in Table 14 that the in-KB accuracy of having this additional nil detection head does not worsen when compared with the models reported in Table 10.

TAC Split	Ours: Multilingual Contextual
En-News	0.826
En-Discussion Forums	0.589
Es-News	0.802
Es-Discussion Forums	0.697
Zh-News	0.671
Zh-Discussion Forums	0.607

Table 13 *F1 results for nil detection on TAC 2016 splits*

TAC Split	Ours: Multilingual Contextual	Ours: Multilingual Contextual, w/ nil detection
En-News	0.9000	0.908
En-Discussion Forums	0.8851	0.891
Es-News	0.9180	0.924
Es-Discussion Forums	0.8986	0.886
Zh-News	0.8409	0.854
Zh-Discussion Forums	0.8176	0.87

Table 14 *In-KB accuracy in a multilingual scenario for original DCA model and our embedding vocabulary – comparison between using nil detection and not using it*

4.3. Story Segmentation

Speaker diarization, the process of distinguishing speakers within an audio segment, is a crucial task in audio analysis that facilitates various downstream applications such as transcription and automated captioning. With the advent of deep learning, segmentation techniques have significantly evolved, yet they often grapple with the intricacies of real-world audio, such as speaker overlap and diverse acoustic environments. We investigated a robust diarization system leveraging the VBx method from Landini et al. (2022), which applies a Bayesian hidden Markov model (HMM) integrated with x-vector clustering to improve speaker segmentation. This system showcases efficiency and precision in processing audio, aiming to surpass previous systems' limitations and setting a novel standard in speaker diarization technology. The final story segmentation of SELMA utilizes the following steps:

- *Speaker Embedding Extraction:* We extract x-vectors from the input audio using a deep neural network. It's powered by a state-of-the-art ResNet-based architecture optimized for capturing the distinct voice characteristics necessary for speaker differentiation. Eventually, these x-vectors serve as dense representations of speaker characteristics, forming the basis for differentiation among speakers.

  The architecture below represents ResNet101 architecture described in Singh and Ganapathy (2021), where the network processes 64 log Mel filter bank features extracted at 10 ms intervals across a 25 ms window. The model utilizes 4-sec-second segments, equating to 400 frames, applying standard ResNet blocks, and statistical pooling to encapsulate mean and standard deviation across time. This is succeeded by a linear transformation stage, which compacts the output to a fixed dimensional vector representation (256-dimension), which is then integral for speaker identification.

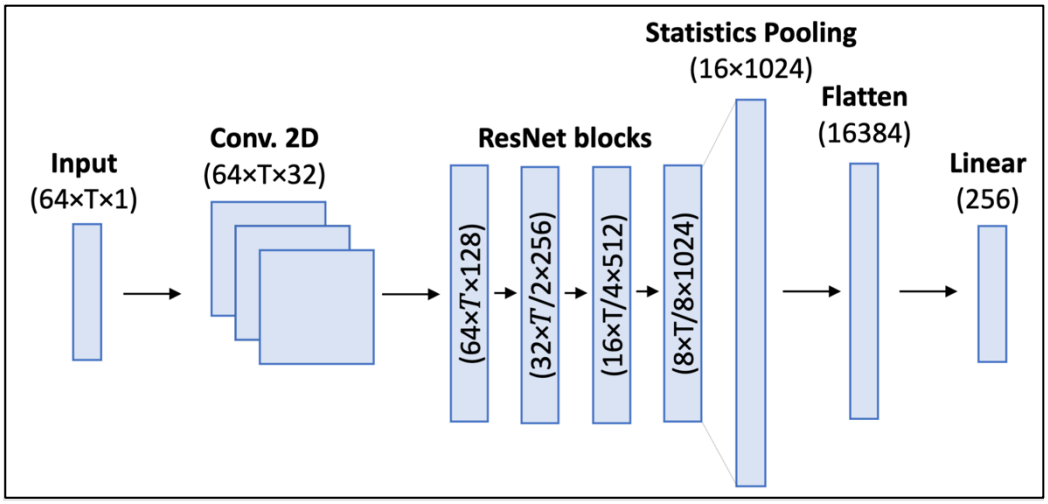

Figure 11 *Architecture of the x-vector where T indicates the number of input frames*

Speaker-specific distributions are then derived via probabilistic linear discriminant analysis (PLDA), which discriminates between speakers by modeling within- and between-speaker variance of x-vectors. Note that the PLDA on such x-vectors is later used to operate on the extraction from much shorter 1.5 s segments. This mismatch, however, does not seem to affect diarization performance negatively.

- *Bayesian HMM for Clustering:* Central to our methodology is the Bayesian HMM, where each state correlates with an individual speaker, and state transitions represent speaker changes. It uses advanced statistical modeling to associate extracted x-vectors with individual speakers via posterior distribution estimation of speaker labels. Typically, the HMMs are initialized with more speakers, and we use this behavior to drop the redundant speakers (i.e., to estimate the number of speakers).

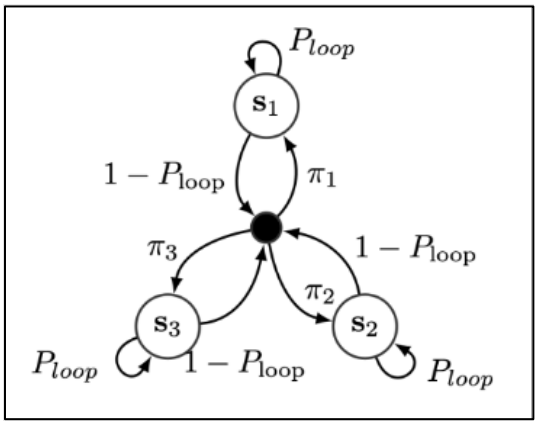

Figure 12 *HMM model for 3 speakers (1 state per speaker), with a non-emitting (initial) state*

In the context of our segmentation, this HMM treats the audio stream as a sequence of observable events generated by transitions between hidden states, which correspond to the different speakers. The speaker (HMM state) specific distributions are derived from a PLDA and each state in the model has its own probability distribution, and the transitions between these states are governed by a set of probabilities as well. The "hidden" aspect comes from the fact that while we observe the data (the audio signal), we do not observe which state (speaker) is responsible for generating each part of the data.

- *Variational Bayes Inference:* The variational Bayes (VB) technique described in Valente et al. (2010) approximates complex integrals for the posterior computation, iterating to align speech segments with corresponding speaker states accurately. Therefore, it adjusts for any overlapping speech and outputs a time-stamped, speaker-tagged transcript of the input audio. VB transforms the challenge of posterior computation, often intractable due to complex integrals, into an optimization problem. It applies a family of simpler distributions and seeks

to find the member of this family that most closely approximates the actual posterior distribution.

The inference process involves iteratively adjusting the parameters of the approximating distribution to minimize the Kullback-Leibler (KL) divergence between the approximate and the accurate posterior distributions. This is achieved by optimizing the Evidence Lower Bound (ELBO), a proxy to the otherwise intractable log marginal likelihood. In SELMA's implementation of speaker diarization, VB inference plays a pivotal role in assigning segments of an audio stream to different speaker states. It manages the uncertainty and variability inherent in real-world broadcast data by iteratively refining these assignments, leading to the identification and labeling of speaker segments even if overlapping speech exists.

For the evaluation, the system exhibited outstanding diarization performance, with a Diarization Error Rate (DER) that surpasses previous benchmarks on standard dataset, VoxConverse from Chung et al. (2020). DER is a comprehensive measure that encapsulates three types of errors: speaker error rate (SER), which quantifies the duration incorrectly attributed to a speaker; false alarm (FA), the time misclassified as speech; and missed speech (Miss), representing speech not attributed to any speaker. The total amount of speech, including overlaps, is the denominator in this calculation. In our system's evaluation, where voice activity detection (VAD) is considered an oracle, FA is negligible, thereby rendering the Miss error as the primary contributor to the DER alongside SER.

As a baseline, we also provide the result of a standalone hierarchical clustering (AHC) of x-vectors, where its threshold is tuned for optimal performance. In comparison to traditional approaches like Kaldi with the Sell et al. (2018) setup, our method demonstrated a significant reduction in both false alarm and miss rates, indicating a notable advancement in the precision of speaker diarization. Our results underscore the effect of overlap handling and the choice of collar size on the DER metric, with the 0.25 collar size yielding better results compared to no collar, and the inclusion of overlap analysis in the computation typically increasing the DER.

Collar	*Overlap*	*System*	*DER [%]*
0.25	No	Kaldi [Sell et al. (2018)]	16.25
		AHC	12.43
		VBx (Ours)	10.31
0	Yes	Kaldi [Sell et al. (2018)]	28.10
		AHC	22.34
		VBx (Ours)	17.76
0.25	Yes	Kaldi [Sell et al. (2018)]	23.49
		AHC	19.55
		VBx (Ours)	13.28

Table 15 *Comparative analysis of speaker segmentation systems over the Diarization Error Rate (DER)*

In summary, SELMA's story segmentation system represents a significant stride forward in audio processing. The implications of our results are far-reaching, suggesting that integrating Bayesian HMM with x-vector clustering in diarization tasks offers a substantial improvement over existing methods. Its precision, efficiency, and integration-friendly design make it an advantageous tool for various applications, from automatic transcription to advanced audio analytics. Our system provides the efficacy of combining deep learning with traditional probabilistic models in speaker segmentation tasks.

4.4. Online News Classification

We compared the results of our new approaches on the News Classification problem to the model previously described in report D5.1 of the SUMMA project, hereafter referred to as "multi-CNN". We report micro-F1 scores of our models, trained on the Lusa Portuguese news dataset with IPTC subject labels. We also report zero-shot cross-lingual results on the smaller English and Spanish datasets. Table 16 shows the results of our sentence embedding attention-based models. We compare the results of using a single query to generate a single representation of the model; Three queries, corresponding to the three depths of the label hierarchy, to develop three representations; And having each label learn its own query. As a baseline, we also present the results of averaging all sentence embeddings in a document and using the resulting vector for classification.

Model	Portuguese F1	English F1	Spanish F1
Multi-CNN	64.33%	49.32%	52.61%
DistilUSE + average	65.08%	54.24%	49.16%
DistilUSE + global attention	66.77%	53.19%	60.05%
DistilUSE + hierarchy depth attention	67.40%	52.13%	61.30%
DistilUSE + label attention	66.48%	54.52%	60.63%

Table 16 *F1 performance of sentence embedding attention-based models on Portuguese, English, and Spanish testing datasets (English and Spanish are zero-shot languages)*

Table 17 shows the results of our AttentionXML based models. We compare the results of using a traditional AttentionXML with a word embedding layer using the multilingual BPEmb embeddings and using a multilingual mBERT model to generate the contextual word embeddings that are fed into the biLSTM of AttentionXML. Our current results and incremental improvements of F1 scores over previous models show the promise of the current direction of work. As future work we intend on leveraging the information in the label descriptions available in the IPTC vocabulary to generate better label embeddings. This is similar to work done in the past by Mittal et al.

Model	Portuguese F1	English F1	Spanish F1
Multi-CNN	64.33%	49.32%	52.61%
AttentionXML + BPEmb	68.63%	33.26%	55.29%
AttentionXML + mBERT	70.10%	52.88%	64.36%

Table 17 *F1 performance of sentence embedding attention-based models on Portuguese, English, and Spanish testing datasets, for models trained on the Lusa dataset (English and Spanish are zero-shot languages)*

A drawback to be tackled is the limited input size of 512 tokens on the AttentionXML+mBERT model. Approaches to this issue include using BERT style models that are pretrained for longer inputs, such as the Longformer (Beltagy et al. 2020). Alternatively, we intend on experimenting with training AttentionXML's biLSTM to join the concatenated outputs of consecutive mBERT forward passes.

As expected of models that are fine-tuned on a monolingual Portuguese dataset, the best results are obtained on the Portuguese language test sets. This suggests that some multilingual performance of the pretrained models is lost in our experimental setup. We have also trained some of the described models on the multilingual dataset created from joining the Finnish and Portuguese news datasets from STT and Lusa, respectively. It should be noted that these results cannot be fairly compared to the ones shown on the previous tables because the label space changed to include labels that were added for being present in the Finnish dataset. The results for these models are shown in Table 18, along with the scores of the old Multi-CNN model, which has not been retrained on the Finnish dataset, and along a hybrid model of AttentionXML with a multilingual Roberta-Large.

Model	Portuguese F1	Finnish F1	English F1	Spanish F1
Multi-CNN*	64.33%	14.93%	49.32%	52.61%
AttentionXML + BPEmb	67.15%	67.86%	29.04%	43.37%
AttentionXML + mBERT	67.69%	66.28%	55.44%	64.04%
AttentionXML + Roberta	70.32%	68.58%	56.26%	62.81%

Table 18 *F1 performance of sentence embedding attention-based models on Portuguese, Finnish, English, and Spanish testing datasets, for models trained on the Lusa+STT dataset (*excluding Multi-CNN) (English and Spanish are zero-shot languages)*

SmartTags: Continuously Learning to Suggest News Articles According to User Preferences

As mentioned in section 3.4, we used an LLM in a few-shot setting to generate tag descriptions for each pair of keywords. The prompt was the following:

> Here are some examples of a keyword pair and the respective description:
>
> Ukraine and Nato: Reactions and updates on the possibility of Ukraine joining NATO, including demonstration of interest by Ukrainian leaders, remarks by world leaders and threats from Russia.
>
> Rising prices and energy: News about the rising prices of energy worldwide, including impacts on consumers, plans from governments to tackle them, discussion of causes, and investment in green energy. Reports focusing on inflation in other sectors are not relevant.
>
> Taiwan and politics: Updates on the political situation of Taiwan, including visits of foreign heads of state and military displays by China in response.
>
> Dengue and South America: Reports on the ongoing Dengue outbreak in South America. This includes death counts, new strains and symptoms. Reports of infections in other continents are also relevant.
>
> Evacuation missions and Sudan: Updates on the ongoing efforts from European countries to evacuate their citizens from Sudan, following the armed conflict. Reports that focus only on the armed conflict itself are not relevant.
>
> Complete the description for the following keyword pair: [first keyword from tag] and [second keyword from tag]:

We show the LLM output for the keyword pair "'Brazil'" and "Dilma Rousseff" as an example:

> Reports on the political situation and news related to Dilma Rousseff, the former President of Brazil. This includes updates on her political career, her relationship with other politicians and the current administration, and any legal proceedings she may be involved in. Reports on other former or current Brazilian politicians are not relevant.

As we can observe in this example, the model successfully mimicked the style of the human-written descriptions, mentioning both information that is and that it is not relevant for the tag.

To train the two SVM models, we used the articles that contained 10 selected keyword pairs in their keyword list, and we used the LLM to generate tag descriptions for each of the 10 keyword pairs. An article - tag description pair was considered positive if the article contained the tag keywords in its keyword list, otherwise it was considered negative. A pair of articles was considered positive if the intersection of their keyword lists was not empty, otherwise it was considered negative.

We evaluated the system on the same tags it saw during training, as well as the more challenging and realistic setting of tags not seen in the training data. Table 19 shows the average test results across the 10 tags seen during training for different score aggregation functions. We report the area under the precision-recall curve (PR-AUC) and the F1 score for the best threshold found in the validation set. These results show small differences across different score aggregation functions, with a small advantage of the Mellowmax in terms of F1 score.

Score Aggregation	F1	PR-AUC
Max	0.453	0.395
Mean	0.468	0.441
Mellowmax	**0.509**	0.440
Probability-based	0.454	**0.442**

Table 19 *Average F1 scores and precision-recall AUCs of the SmartTags system on the test sets of the 10 tags seen during training*

In the more challenging scenario of tags unseen during training, we evaluated the system performance in a realistic few-shot setting, where the number of positive examples is relatively small (50 or less). The results are displayed in Figure 13 for three different tags and where the Mellowmax aggregation function was used. The first observation is that the performance of the system with only one positive example is generally low, which indicates that the tag descriptions alone are not sufficient for the model to make accurate predictions. As expected, the performance improves as the number of positive examples increases, but the plateau is low and reached with a small number of positive examples. Therefore, the overall quality of the model is still insufficient for solving the task at hand, and further research endeavors should be undertaken in that direction.

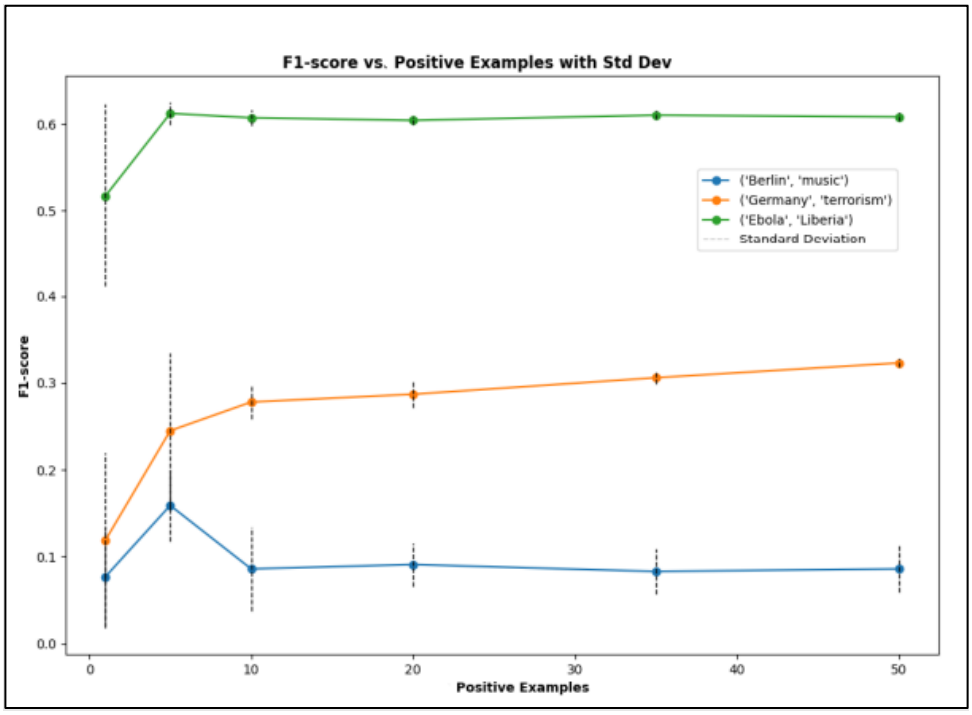

Figure 13 *F1 scores of the SmartTags system on three novel tags not encountered during training, as influenced by the number of positive examples provided*

4.5. Online News Clustering

We follow previous work on this task and evaluate our system on a news clustering dataset (Rupnik et al. (2016)). Besides the three main languages (English, Spanish, and German), this dataset also provides a significant number of documents in Chinese and Russian, as well as documents in Slovenian, Croatian, French, and Italian.

Systems	BCubed			Standard			Clusters
	F1	*P*	*R*	*F1*	*P*	*R*	
Miranda et al. (2018)	-	-	-	84.00	83.0	85.00	-
Linger et al. (2020)	82.06	80.25	83.97	86.49	85.11	87.92	606
4-F Rank + Accept.	88.02	91.31	84.95	92.34	97.26	87.09	957
8-F Rank + Accept.	89.24	92.62	86.11	93.76	97.66	90.15	1023
8-F Rank + Accept. + Merge	90.10	89.70	90.51	97.21	97.01	97.42	812

Table 20 *Cross-lingual clustering performances on the news clustering test dataset where P and R represent the precision and recall respectively*

The samples allow us to roughly preview the system's performance in other languages besides the ones it was trained in. The dataset is composed of 34,687 news documents, and it is divided into two sets: a training set comprised of 20,813 articles and a test set that contains 13,874 articles. For cross-lingual clustering, as shown in Table 20 our system achieves state-of-the-art performance on BCubed F1 (Amigó et al. (2009)) (+8.04) and the standard F1 (+11.33) despite producing a larger number of clusters. We also perform an ablation study that shows the relative importance of system components. 4-F Rank+Accept. refers to the clustering system with a 4-feature ranking and acceptance model. Adding the other features, such as 8-F Rank+Accept., improved both standard (+1.42) and BCubed F1 (+1.22). Finally, the cluster merge model is added to our system, which results in gains for both standard (+3.35) and BCubed F1 (+0.86).

Languages	BCubed			Standard			Clusters
	F1	*P*	*R*	*F1*	*P*	*R*	
Chinese	96.18	100.00	92.65	99.07	100.00	98.16	28
Slovenian	76.92	100.00	62.50	79.67	100.00	66.21	12
Croatian	77.85	100.00	63.73	74.99	100.00	60.00	5
French	98.50	100.00	97.04	99.69	100.00	99.39	3
Russian	100.00	100.00	100.00	100.00	100.00	100.00	1
Italian	98.86	100.00	97.75	98.78	100.00	97.59	3

Table 21 *Clustering performances on other languages where P and R represent the precision and recall respectively*

Given the nature of our system, we evaluated it on the remaining languages of the dataset, as shown in Table 21. Our ranking, acceptance, and cluster merge models were not trained on any data from these languages (except for Chinese), making this a zero-shot clustering scenario. Chinese, French, Russian, and Italian document clustering had high F1 scores, with results above 95%, and both Slovenian and Croatian had initial clustering scores above 70%.

Regarding future work, a relevant approach to follow is the implementation of high-performance vector search in order to improve clustering speed and scalability, which takes advantage of the current fully dense clustering space. Taking the feedback of users into account on the clustering process in order to fine-tune the models is also a pertinent direction. Regarding the improvement of the current evaluation scores, following the work on entity-aware contextual embeddings is also a relevant approach, with the main obstacle being the need of said entity-awareness to cover all of the SELMA languages.

4.6. News Summarization

Monolingual Text Summarization

We evaluate our energy-based re-ranking model (EBR) described in Section 3.6 against a baseline BART system with the usual beam search decoding algorithm and against other improved summarization systems, namely: BRIO (Liu et al., 2022), which employs a ranking loss as an additional term on the training of the abstractive system; CLIFF (Cao & Wang, 2021), which uses data augmentation techniques and contrastive learning to enhance the factual consistency of the summaries; DAE (Goyal & Durret, 2021), which detects and discards non-factual tokens from the training data; FASum (Zhu et al., 2021), which incorporates knowledge graphs also to enhance factual consistency; and SumRerank (Ravaut et al., 2022), which employs a mixture of experts to train a re-ranker on the combination of various metrics. For our model and for SumRerank, we sample 8 candidate summaries from BART using diverse beam search (Vijayakumar et al., 2016). The models are evaluated on two benchmark datasets for abstractive summarization: CNN/DailyMail (Hermann et al., 2015) and XSum (Narayan et al., 2018), both containing news articles paired with their respective reference summaries. In XSum each summary consists of a single sentence, while in CNN/DailyMail it can comprise three sentences or more. Regarding the automatic evaluation metrics, apart from the usual ROUGE scores, we also measured the QuestEval (Scialom et al., 2021) and CTC scores (Deng et al., 2021), which are transformer-based metrics that exhibit a stronger correlation with human judgment. The results of the baselines and of our EBR trained with the CTC metric are in Table 22.

Models	CNN/DailyMail				XSum			
	R2	***QE***	***Cons***	***Rel***	***R2***	***QE***	***Cons***	***Rel***
BART	20.75	43.28	95.01	61.75	19.42	28.27	83.18	52.23
BRIO	**24.06**	43.49	89.61	60.75	-	-	-	-
CLIFF	20.88	43.28	94.68	60.38	**21.41**	**29.34**	82.57	51.92
DAE	-	-	-	-	14.19	29.20	79.45	51.05
FASum	17.68	42.87	94.30	57.91	9.97	24.35	75.45	39.42
SumRerank	21.73	43.61	95.07	62.49	21.40	28.76	83.00	52.75
EBR [Ours]	20.87	**43.79**	**96.15**	**63.32**	19.72	28.66	**86.03**	**54.74**

Table 22 *Results of our model and baselines on each of the automatic evaluation metrics. (R2: ROUGE-2, QE: QuestEval, Cons: CTC consistency, Rel: CTC relevance)*

We see that our model outperforms or is competitive with the remaining in all the metrics except ROUGE, which is known to correlate poorly with human judgment. Interestingly, despite the fact that our model was trained with the CTC scores only, it yields improvements over BART in ROUGE and QuestEval metrics as well.

Even though the results of automatic evaluation are promising, directly optimizing for a metric is risky as none of these metrics correlate perfectly with human judgment. For this reason, it is crucial to conduct a human evaluation. Specifically, we asked the judges to make pairwise comparisons between the summaries generated by three models: BART, CLIFF, which was the strongest published baseline at the time we conducted this study, and our EBR trained with the CTC scores. For each source document, we presented three pairs of summaries consecutively, which correspond to all the pairwise combinations of the summaries generated by the three systems. Then, we asked the judges to rank the summaries in each pair according to three criteria: factual consistency, relevance, and fluency. For each criterion, the judges had to evaluate whether the first summary was better than, tied with, or worse than the second. We randomly sampled 30 source documents from the test set of CNN/DailyMail and another 30 from the test set of XSum, so each judge was asked to compare 180

pairs of summaries. The results are presented in Table 23. The first observation is that our EBR model succeeds at improving the quality of the candidates sampled from BART on the CNN/DailyMail dataset in all three criteria. On XSum, the improvements are marginal or even absent, except on the fluency dimension. Surprisingly, the comparison of our model with CLIFF contradicts the results of the automatic evaluation (Table 22), especially on the XSum dataset. Further analysis conducted in our work shows that the primary cause for this contradiction are flaws in the CTC metrics that our model was trained to mimic. Specifically, the CTC consistency metric often fails at detecting factual inconsistencies, especially when the summaries are highly abstractive as is the case in XSum.

Despite the improvements obtained by our approach, the lack of reliable metrics to automatically assess summary quality, particularly its factual consistency, spoils its effectiveness in more abstractive settings. We reemphasize the difficulty of evaluating summary quality automatically and therefore this is a topic that should deserve our attention in future work. Moreover, most of the aforementioned transformer-based metrics (e.g., CTC scores and QuestEval) are only available for English and therefore the applicability of our method to non-English data is not straightforward.

Models	CNN/DailyMail			XSum		
	FC	***R***	***F***	***FC***	***R***	***F***
CLIFF is better	.17	.33	**.33**	**.25**	**.32**	**.27**
Tie	.65	.24	.40	.63	.63	.68
BART is better	**.18**	**.43**	.27	.12	.05	.05
EBR is better	**.13**	**.30**	**.24**	**.15**	.12	**.30**
Tie	.80	.52	.58	.72	.77	.63
BART is better	.07	.18	.18	.13	.12	.07
EBR is better	.12	**.45**	**.32**	.10	.08	.07
Tie	.68	.20	.42	.63	.63	.88
CLIFF is better	**.20**	.35	.27	**.27**	**.28**	**.08**
Agreement	.50	.63	.54	.56	.58	.87
Strong disag.	.01	.11	.08	.01	.00	.00

Table 23 *Proportion of times that each model was considered the best for the human judges in each pairwise comparison according to three criteria: factual consistency (FC), relevance (R), and fluency (F). Rows "Agreement" and "Strong disag." show, respectively, the proportion of times that the two judges agreed and chose opposite options on the pairwise comparisons*

Cross-Lingual Text Summarization

We utilized the CrossSum dataset (Bhattacharjee et al., 2023) in our experiments. This dataset contains pairs of documents in one language along with summaries in various languages, spanning over 1500 language pairs. Our focus lies in the multi-target setting, where a single document undergoes summarization in multiple languages. To achieve this, we organized the CrossSum data into clusters, where each cluster comprises different language versions of the same document along with their corresponding summaries. The following languages were used in our experiments: *en, fr, es, pt, ar, ru,* and *zh* (simplified).

Besides evaluating the relevance of the produced summaries by comparing them with the corresponding references, we also want to evaluate the extent to which summaries in different languages of the same document contain identical information. For the former, we compute the usual ROUGE scores with respect to the reference summaries. For the latter, we use COMET (Rei et al., 2020), a reference-free metric for machine translation. Specifically, we evaluate the COMET scores of the produced summaries in each target language relative to the summary generated in the source language.

As explained in section 2.6, our approach involves sampling multiple candidate summaries for each target language and then using a multilingual encoder to choose the candidate most similar to the summary generated for the source language. This approach is compared with two other standard approaches: i) independently generating summaries for each target via beam search, and ii) generating a summary in the source language and subsequently employing a machine translation model to translate it into each target language. In the reported experiments, we use English as the source language.

We used an mT5 base model (Xue et al., 2021) fine-tuned in the CrossSum dataset as the cross-lingual summarization model, NLLB-200-1.3B (Costa-jussà et al., 2022) as the MT model, and SONAR (Duquenne et al., 2023) to obtain multilingual sentence embeddings. The results for multi-target summarization from English are shown in the table below.

Method	**ROUGE-2**							**COMET**						
	ar	***es***	***en***	***fr***	***pt***	***ru***	***zh***	***ar***	***es***	***en***	***fr***	***pt***	***ru***	***zh***
Beam Search	8.2	12.7	16.2	19.1	10.6	8.6	16.0	55.6	60.5	-	58.2	59.7	63.2	60.7
Re-ranking with pivot	8.4	12.8	16.2	17.1	10.9	8.5	15.1	58.9	64.3	-	64.2	62.8	66.9	65.4
Re-ranking pivot-free	8.7	12.7	15.9	17.1	10.8	8.5	15.2	59.8	64.5	-	62.9	63.0	67.7	65.3
Summ+Transl.	7.3	11.0	16.2	16.6	10.4	7.4	7.1	83.9	88.1	-	88.6	87.1	87.8	80.3

Table 24 *Results of multi-target cross-lingual summarization. English is used as the source language in all cases*

Our main finding is that our re-ranking method consistently outperforms beam search in terms of COMET scores in all scenarios, while maintaining ROUGE scores. This shows that it effectively enhances the semantic coherence of the generated summaries across various target languages without compromising similarity to the references. The process of monolingual summarization followed by machine translation ensures the highest consistency across different target languages, as demonstrated by the COMET scores. However, this approach results in a significant reduction in ROUGE scores across multiple configurations.

In our upcoming publication, we will extend these experiments to other source languages besides English. Additionally, we will provide results for an experiment where we use a large language model in a zero-shot setting as a replacement for the mT5-base summarizer.

Multilingual Multi-Document Summarization

We compare our approaches with the centroid-based methods from Ghalandari (2017), and Lamsiyah et al. (2021). To be consistent with the remaining methods, the approach by Ghalandari (2017), was implemented on top of contextual sentence embeddings instead of TF-IDF. Additionally, we perform ablation evaluations in three scenarios: i) a scenario (BS) where we do not use the centroid estimation model and rely solely on the beam search for the sentence selection step; ii) a scenario (BS+GS) identical to the previous one, except that we perform the greedy search step after the beam search; iii) two scenarios (CeRAI and CeRA) where we utilize the centroid estimation model with and without incorporating interpolation, and apply the BS+GS algorithm on the predicted centroid.

We used four English datasets, Multi-News, WCEP-10, TAC2008, and DUC2004, and one multilingual dataset, CrossSum, in our experiments. CrossSum was conceived for single-document cross-lingual summarization, so we had to adapt it for multilingual MDS. This adaptation results in clusters that encompass documents in multiple languages, with each cluster being associated with a single reference summary containing sentences in various languages.

We used the centroid-estimation models trained on Multi-News to evaluate CeRA and CeRAI on WCEP-10, TAC2008, and DUC2004 since these datasets do not provide training splits. For CrossSum, the languages present in the test were unseen during the training phase. We present the ROUGE-2 recall scores of our model and compared methods in the table below.

Method	*Multi-News*	*WCEP-10*	*TAC 2008*	*DUC 2004*	*Cross Sum*
Ghalandari (2017)	16.07	15.09	7.36	6.82	10.03
Lamsiyah et al. (2021)	13.92	16.10	7.91	**7.80**	10.45
BS (Ours)	16.22	15.64	8.10	7.03	10.16
BS+GS (Ours)	16.70	16.41	8.16	7.46	10.85
CeRA (Ours)	17.98	**17.46**	8.27	7.31	11.67
CeRAI (Ours)	**17.99**	17.24	**8.37**	7.72	**11.73**

Table 25 *ROUGE-2 recall results of different extractive methods on the considered test sets.*

The first observation is that BS alone outperforms Ghalandari, 2017, in all datasets, with additional improvements obtained when the greedy search step is also performed (BS+GD). This was expected since our approach explores the candidate space more thoroughly. The two methods using the centroid estimation model (CeRA and CeRAI) improve R2-R significantly in Multi-News and WCEP-10 and perform at least on par with Lamsiyah et al., 2021, in TAC2008 and DUC2004. It is also worth noting that CeRA and CeRAI were only trained on the Multi-News training set and nevertheless performed better or on par with the remaining baselines on the test sets of the remaining corpora. Incorporating the interpolation step (CeRAI) appears to yield supplementary enhancements compared to the non-interpolated version (CeRA) across various settings, which we attribute to this method adding regularization to the estimation process, improving results on harder scenarios.

In the multilingual dataset CrossSum, we again observe the superiority of the centroid estimation models, CeRA and CeRAI, in comparison to all the remaining methods. Most notably, these models prove to be useful even when tested with languages unseen during the training phase, underscoring their robustness and applicability in a zero-shot setting. Further results, experimental details, and an extended discussion can be found in Gonçalves et al. (2023).

Speech-to-Text Summarization

For this work, we built a dataset for speech-to-text (S2T) abstractive summarization of broadcast news in French, that was built from articles that can be found in the EuroNews website. Each news article from EuroNews has an audio, an abstractive summary of the news content and the article body. Since the latter is not always a perfect transcript of the audio, we employed an automatic procedure for selecting the news articles whose article bodies are perfect (or almost perfect) transcripts of the audios. An XLSR-based ASR model was used to produce artificial transcripts from the audios. Afterwards, the word error rate (WER) evaluation metric was applied between the automatically generated transcript and the article body. A threshold for the WER of 45% was set, such that articles associated with higher values of WER were discarded. The remaining articles were randomly shuffled and separated into three distinct splits with sizes of 13380, 1672 and 1673 for the train, validation and test splits, respectively. The mean audio duration per article is about 87s.

Our end-to-end approach (E2E) is contrasted with two cascaded methods: one comprising automatic speech recognition (ASR) followed by abstractive summarization, and another utilizing ASR followed by extractive summarization. For automatically assessing the performance of the different implementations developed in this work, we make use of the ROUGE metrics. The decoding for the cascade and end-to-end abstractive summarizers is performed with beam search. Table 26 compares the ROUGE scores for the extractive baseline and both cascade and E2E abstractive summarizers on the test split of our dataset. We also performed ablation studies for the following cases: the S2T abstractive summarizer is not fine-tuned on the summarization data after the pre-training of the cross-modal adapter (nFT); there is no fine-tuning and the cross-modal adapter additionally does not make use of its predictions for the end-of-sequence positions of the sequences of textual embeddings and uses instead the gold ones (G-EOS); the pre-training of cross-modal adapter is not performed and the S2T abstractive summarizer is directly trained using the summarization data.

Method	*R-1*	*R-2*	*R-L*	*R-Lsum*
Cascade	41.6	26.2	35.7	37.6
E2E	37.8	23.7	32.8	33.9
E2E (nFT)	30.0	16.1	25.9	26.6
E2E(G-EOS)	29.8	15.9	25.7	26.5
E2E (nPre)	16.8	2.4	12.6	13.2
Extractive	23.8	8.3	17.9	18.8

Table 26 *ROUGE scores of the evaluated approaches.*

All the abstractive systems outperform the extractive baseline, which was expected given that the target summaries from our corpus are abstractive. The E2E model performs worse than the abstractive cascade model, as measured by ROUGE scores. This contrasts with the fact that, theoretically, E2E modeling allows leveraging non-verbal and acoustic information besides the linguistic one from transcripts, which is the only type of information that cascade systems have access to. Regarding the ablation studies, by comparing the performance of the E2E and E2E (nFT) models, it is found that fine-tuning the S2T abstractive summarizer after the pre-training of the cross-modal adapter significantly improves the ROUGE scores with a relative increase on the interval of 25%-50%. The similarity between the ROUGE scores of the E2E (nFT) and E2E (G-EOS) models allows us to conclude that the cross-modal adapter performs equally well either when using its own predictions for the end-of-sequence positions of the sequences of textual embeddings or when using the ground truth ones. Finally, the gap between E2E and E2E (nPre) proves that the proposed pre-training of the cross-modal adapter provides a very significant performance increase.

These results show that the E2E abstractive summarizer underperforms with respect to the cascade one. This under-performance may be explained if one considers the several sub-modules of the cascade and E2E summarizers. Both make use of a W2V2-based model either for speech recognition or plain speech feature extraction. The T2T abstractive summarizer of the cascade system and the S2T abstractive summarizer of the E2E system share the same decoder but differ strongly on the

encoder. Thus, the limited performance of the proposed novel E2E implementation when compared with the cascade system must be sourced on the realization of the cross-modal adapter. We have strong reasons to believe that the large text-to-text summarization corpus, to which the encoder of the text-to-text summarizer was exposed during its training for abstractive summarization, played a significant role. It is likely that this enormous amount of external data makes the text encoder generate much richer textual latent representations than the ones the cross-modal adapter could possibly generate, given that it only had access to the summarization training data from our dataset during its development.

Further implementation details and experimental results, including a human evaluation study, can be found in the publication by Monteiro and Pernes (2023).

5. Conclusions

In this report, we present an overview of the research and development undertaken in the SELMA work package two, WP2. In particular, we present the advances achieved in named entity recognition, entity linking, story segmentation, news summarization, online news classification, and clustering. Significant progress was made on the multilingual NER achieving very impressive zero shot results and allowing us to use only one model for all tested languages. The work done on multilingual and cross-lingual summarization is novel and stresses the commitment of SELMA in developing truly multilingual models that can break language barriers. Our improvements on the explainability of the classification models will enable the use of these models in other scenarios where human supervision is essential. Our work on clustering and summarization was accepted at the SIGIR conference and EMNLP workshops.

All the different components developed in WP2 are the results of our research effort to find the systems that better suit the use-cases of SELMA. These components were integrated on UC1 and UC2 as described in D2.5 and D2.6.

Bibliography

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems,* 33, 12449-12460.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-document Transformer. *ArXiv:2004.05150.*

Bhattacharjee, A., Hasan, T. Ahmad, W. U., Li, Y. F. Kang, Y. B., & Shahriyar, R. (2023). CrossSum: Beyond English-Centric Cross-Lingual Summarization for 1,500+ Language Pairs. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Toronto, Canada. Association for Computational Linguistics (pp. 2541–2564,).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics.*

Cao, S., & Wang, L. (2021). CLIFF: Contrastive Learning for Improving Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 6633-6649).

Cardoso, R., Marinho, Z., Mendes, A., & Miranda, S. (2021). Priberam at MESINESP Multi-label Classification of Medical Texts Task. *Proceedings of International Conference of the CLEF Association.*

Chalkidis, I., Fergadiotis, M., Kotitsas, S., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). An Empirical Study on Large-scale Multi-label Text Classification Including Few and Zero-shot Labels. *Empirical Methods in Natural Language Processing (EMNLP).*

Chen, M., Duquenne, P. A., Andrews, P., Kao, J., Mourachko, A. Schwenk, H., & Costa-Jussà, M. R.. 2023. BLASER: A Text-Free Speech-to-Speech Translation Evaluation Metric. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9064–9079, Toronto, Canada. Association for Computational Linguistics.

Chung, J. S., Nagrani, A., & Zisserman, A. (2018). Voxceleb2: Deep Speaker Recognition. *ArXiv:1806.05622.*

Chung, J. S., Nagrani, A., Coto, E., Xie, W., McLaren, M., Reynolds, D. A., & Zisserman, A. (2019). VoxSRC 2019: The First VoxCeleb Speaker Recognition Challenge. *ArXiv:1912.02522.*

Chung, J. S., Huh, J., Nagrani, A., Afouras, T., & Zisserman, A. (2020). *Spot the conversation: Speaker diarisation in the wild. arXiv preprint arXiv:2007.01216.*

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672.*

De Cao, N., Aziz, W., & Titov, I. (2021). Highly Parallel Autoregressive Entity Linking with Discriminative Correction. *Empirical Methods in Natural Language Processing (EMNLP).*

Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D., & Dehak, R. (2011). Language Recognition via i-Vectors and Dimensionality Reduction. *In INTERSPEECH.*

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Deng, M., Tan, B., Liu, Z., Xing, E. P., & Hu, Z. (2021). Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation. *ArXiv:2109.06379.*

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN-Based Speaker Verification. *ArXiv:2005.07143.*

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Dorr, B., Zajic, D., & Schwartz, R. (2003). Hedge Trimmer: A Parse-and-trim Approach to Headline Generation. *Proceedings of the HLT-NAACL on Text Summarization Workshop.*

Duquenne, P. A., Schwenk, H., & Sagot, B. (2023). Sentence-level multimodal and language-agnostic representations. arXiv preprint arXiv:2308.11466.

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.

Ferreira, P., Cardoso, R., & Mendes, A. (2021). Priberam Labs at the 3rd. Shared Task on SlavNER. *Proceedings of the Balto-Slavic Natural Language Processing Workshop.*

Ganea, O.-E., & Hofmann, T. (2017). Deep Joint Entity Disambiguation with Local Neural Attention. *ArXiv:1704.04920.*

Gao, S., Cheng, M. M., Zhao, K., Zhang, X. Y., Yang, M. H., & Torr, P. H. (2019a). Res2Net: A new Multi-scale Backbone Architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Gao, Z., Song, Y., McLoughlin, I. V., Li, P., Jiang, Y., & Dai, L. R. (2019b). Improving Aggregation and Loss Function for Better Embedding Learning in End-to-End Speaker Verification System. *In INTERSPEECH.*

Garcia-Romero, D., Snyder, D., Sell, G., McCree, A., Povey, D., & Khudanpur, S. (2019). X-Vector DNN Refinement with Full-Length Recordings for Speaker Recognition*. In INTERSPEECH.*

Garcia-Romero, D., McCree, A., Snyder, D., & Sell, G. (2020). JHU-HLTCOE System for the VoxSRC Speaker Recognition Challenge. *International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Ghalandari, D. G. (2017). Revisiting the Centroid-based Method: A Strong Baseline for Multi-Document Summarization. *Proceedings of the Workshop on New Frontiers in Summarization,* Copenhagen, Denmark. Association for Computational Linguistics (pp. 85-90).

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).

Goyal, T., & Durrett, G. (2021). Annotating and Modeling Fine-grained Factuality in Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1449-1462).

Guo, H., Pasunuru, R., & Bansal, M. (2018). Soft Layer-specific Multi-task Summarization with Entailment and Question Generation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Hasan, T., Bhattacharjee, A., Islam, M. S., Mubasshir, K., Li, Y. F., Kang, Y. B., ... & Shahriyar, R. (2021). XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 4693-4703).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition.*

Heinzerling, B., & Strube, M. (2018). BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. *ArXiv:1710.02187.*

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1* (pp. 1693-1701).

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning* (pp. 2790-2799). PMLR.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation networks. *IEEE Conference on Computer Vision and Pattern Recognition.*

Ioffe, S. (2006). Probabilistic Linear Discriminant Analysis. *European Conference on Computer Vision.*

Johnson, A., Pollard, T., Shen, L., Lehman, L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L., Mark, R., (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data.*

Lamsiyah, S., El Mahdaouy, A., Espinasse, B., & Ouatik, S. E. A. (2021). An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings. *Expert Systems with Applications,* 167, 114152.

Monteiro, R., & Pernes, D. (2023). Towards End-to-End Speech-to-Text Summarization. *Text, Speech, and Dialogue: 26th International Conference, TSD 2023*, Pilsen, Czech Republic, September 4–6, 2023, Proceedings. Springer-Verlag, Berlin, Heidelberg (pp. 304–316).

Kim, Jin-Dong & Ohta, Tomoko & Tateisi, Yuka & Tsujii, Jun'ichi. (2003). GENIA corpus—A semantically annotated corpus for bio-textmining. Bioinformatics (Oxford, England). 19 Suppl 1. i180-2. 10.1093/bioinformatics/btg1023.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., & Lewis, M. (2020). Generalization through Memorization: Nearest Neighbor Language Models. *ArXiv:1911.00172.*

Khandelwal, U., Fan, A., Jurafsky, D., Zettlemoyer, L., & Lewis, M. (2021). Nearest Neighbor Machine Translation. *ArXiv:2010.00710.*

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017). A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019a). Neural Text Summarization: A Critical Evaluation. *ArXiv:1908.08960.*

Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2019b). Evaluating the Factual Consistency of Abstractive Text Summarization. *ArXiv:1910.12840*

Kupiec, J., Pedersen, J., & Chen, F. (1995). A Trainable Document Summarizer. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

Landini, F., Profant, J., Diez, M., & Burget, L. (2022). Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks. *Computer Speech & Language, 71, 101254.*

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the Annual Meeting of Association for Computational Linguistics.*

Linger, M., & Hajaiej, M. (2020). Batch Clustering for Multilingual News Streaming. *Proceedings of the Text2Story Workshop.*

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... & Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics,* 8, 726-742.

Lu, L., Liu, L., Hussain, M. J., & Liu, Y. (2017). I Sense You by Breath: Speaker Recognition via Breath Biometrics. *IEEE Transactions on Dependable and Secure Computing.*

Liu, Y., Liu, P., Radev, D., & Neubig, G. (2022). BRIO: Bringing Order to Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2890-2903).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*

Marinho, Z., Mendes, A., Miranda, S., & Nogueira, D. (2019). Hierarchical Nested Named Entity Recognition. *Proceedings of the Clinical Natural Language Processing Workshop.*

Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations (ICLR).*

Mendes, T. A. A. L. (2023). SmartTags: Continuously learning to suggest news articles according to user preferences. Unpublished master's thesis. Faculdade de Engenharia da Universidade do Porto, Portugal.

Miranda, S., Znotins, A., Cohen, S.B., & Barzdins, G. (2018). Multilingual Clustering of Streaming News. *Empirical Methods in Natural Language Processing (EMNLP).*

Mittal, A., Dahiya, K., Agrawal, S., Saini, D., Agarwal, S., Kar, P., & Varma, M. (2021). DECAF: Deep Extreme Classification with Label Features. Proceedings of the 14th ACM International Conference on Web Search and Data Mining.

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A Large-scale Speaker Identification Dataset. ArXiv:1706.08612.

Nagrani, A., Chung, J. S., Xie, W., & Zisserman, A. (2020). VoxCeleb: Large-scale Speaker Verification in the Wild. Computer Speech & Language.

Nallapati, R., Zhai, F., & Zhou, B. (2017). SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. *AAAI Conference on Artificial Intelligence.*

Narayan, S., Cohen, S. B., & Lapata, M. (2018). Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1797-1807).

Nguyen, K., & Daumé III, H. (2019). Global Voices: Crossing Borders in Automatic News Summarization. *Proceedings of the 2nd Workshop on New Frontiers in Summarization* (pp. 90-97).

Okabe, K., Koshinaka, T., & Shinoda, K. (2018). Attentive Statistics Pooling for Deep Speaker Embedding. *ArXiv:1803.10963.*

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *ArXiv:1904.08779.*

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)*.

Piskorski, J., Babych, B., Kancheva, Z., Kanishcheva, O., Lebedeva, M.Y., Marcinczuk, M., Nakov, P., Osenova, P., Pivovarova, L., Pollak, S., Pribán, P., Radev, I., Robnik-Sikonja, M., Starko, V., Steinberger, J., & Yangarber, R. (2021). Slav-NER: the 3rd Cross-lingual Challenge on Recognition,

Normalization, Classification, and Linking of Named Entities across Slavic Languages. *Proceedings of the Balto-Slavic Natural Language Processing Workshop.*

Ravaut, M., Joty, S., & Chen, N. (2022). SummaReranker: A Multi-Task Mixture-of-Experts Re-ranking Framework for Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4504-4524).

Rei, R., Stewart, C., Farinha, A., and Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In Proceedings of the *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the Empirical Methods in Natural Language Processing.*

Rosales-Méndez, H., Hogan, A., & Poblete, B. (2018). VoxEL: A Benchmark Dataset for Multilingual Entity Linking. *International Semantic Web Conference.*

Rupnik, J., Muhic, A., Leban, G., Skraba, P., Fortuna, B., & Grobelnik, M. (2016). News Across Languages—Cross-lingual Document Similarity and Event Tracking. *Journal of Artificial Intelligence Research.*

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)

Scialom, T., Dray, P. A., Lamprier, S., Piwowarski, B., Staiano, J., Wang, A., & Gallinari, P. (2021). QuestEval: Summarization Asks for Fact-based Evaluation. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 6594-6604).

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the Point: Summarization with Pointer-generator Networks. *Proceedings of Annual Meeting of the Association for Computational Linguistics.*

Sell, G., Snyder, D., McCree, A., Garcia-Romero, D., Villalba, & Khudanpur, S. (2018). Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge. In Interspeech.

Singh, P., & Ganapathy, S. (2021). Self-supervised metric learning with graph clustering for speaker diarization*. In IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).*

Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., & Khudanpur, S. (2018a). Spoken Language Recognition Using X-Vectors. *In Odyssey.*

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018b). X-Vectors: Robust DNN Embeddings for Speaker Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL.*

Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., & Wang, R. (2017). Speaker Identification Features Extraction Methods: A Systematic Review. *Expert Systems with Applications.*

Valente, F., Motlicek, P., & Vijayasenan, D. (2010). Variational Bayesian speaker diarization of meeting recordings. *In IEEE International Conference on Acoustics, Speech and Signal Processing.*

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., & Batra, D. (2016). Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Wang, Q., Downey, C., Wan, L., Mansfield, P. A., & Moreno, I. L. (2018). Speaker Diarization with LSTM. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 483-498).

Yang, X., Gu, X., Lin, S., Tang, S., Zhuang, Y., Wu, F., Chen, Z., Hu, G., & Ren, X. (2019). Learning Dynamic Context Augmentation for Global Entity Linking. *Empirical Methods in NLP and International Joint Conference on NLP (EMNLP-IJCNLP).*

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*.

Yu, J., Bohnet, B., & Poesio, M. (2020). Named Entity Recognition as Dependency Parsing. *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Zeinali, H., Wang, S., Silnova, A., Matějka, P., & Plchot, O. (2019). BUT System Description to VoxCeleb Speaker Recognition Challenge 2019. *ArXiv:1910.12592.*

Zhu, C., Hinthorn, W., Xu, R., Zeng, Q., Zeng, M., Huang, X., & Jiang, M. (2021). Enhancing Factual Consistency of Abstractive Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 718-733).