Research and Innovation Action (RIA) H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu

D3.5 Intermediate Release of Transcription, Punctuation and Translation, Voice Synthesis Capabilities

Work Package	3
Responsible Partner	LIA
Author	Tugtekin Turan, Salima Mdhaffar
Contributors	
Reviewer	Yannick Estève
Version	0.5
Contractual Date	31 December 2022
Delivery Date	22 December 2022
Dissemination Level	Public

Version History

Version	Date	Description
0.1	03/11/2022	Initial Table of Contents (ToC)
0.2	05/12/2022	First Draft
0.3	15/12/2022	Merging the Partner Inputs
0.4	19/12/2022	Internal Review Version
0.5	21/12/2022	Finalization
1.0	22/12/2022	Publishable Version

Executive Summary

The latest version of this SELMA software introduces several new and improved features in the field of natural language processing, with a focus on speech processing under the WP3. This release includes end-to-end neural architectures designed to reduce error propagation and new transfer learning solutions that allow knowledge from one spoken language to be transferred to another.

The SELMA software also takes advantage of automatic to process large amounts of data in different languages to improve the accuracy of the outputs. In the future, the use of self-learning training systems will help to continually improve the performance of WP3 tasks over time.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

In this report, we give the details of our models and software built by the SELMA partners to achieve advanced the state-of-the-art performance. This deliverable addresses the following topics:

- Automatic Speech Recognition (ASR)
- Speech Translation (ST)
- Named Entity Recognition from Speech (NER-S)
- Text-to-Speech Synthesis (TTS)
- Punctuation and Capitalization Recovery (PCR)

The intermediate version of the SELMA release introduces several new and improved features in the fields of Automatic Speech Recognition (ASR), Speech Translation (ST), Named Entity Recognition from Speech (NER-S), Text-to-Speech Synthesis (TTS), and Punctuation and Capitalization Recovery (PCR). These technologies play a crucial role in media monitoring and analysis, and the enhancements in this release make the SELMA software even more powerful and effective.

The ASR module has been improved to provide more accuracy and with an increased number of languages, making it easier to extract valuable information from speech content. The ST module supports more languages and offers improved translation quality via multilingual self-supervised pre-training with unlabeled speech data. The NER-S module has been enhanced to recognize a wider range of named entities from speech, providing valuable insights and context.

The TTS module has been updated to generate more natural and emotional speech, and the PCR module has been improved to recover punctuation and capitalization in the transcribed text over multiple languages by utilizing large language models. These and other enhancements include performance improvements, making the software more stable and efficient, which yield valuable tools for media monitoring staff and journalists. All components are deployed as containers and will be available at our Docker hub, https://hub.docker.com/orgs/selmaproject.

2. Released Components

2.1 Foundation Blocks for Speech Processing: wav2vec 2.0 Models

Speech processing models based on self-supervised learning (SSL) are popular nowadays because they allow us to develop with a smaller amount of annotated data. They can thus be leveraged for many, if not all, the target tasks of the SELMA project as a speech processing block.

During this project, we intend to not only apply these models to our targeted tasks but also to extensively investigate the impact caused by having these processing blocks integrated into different tasks. The table below provides an overview of the wav2vec 2.0 models trained in the context of the first year of the SELMA project.

Available Models					
	Model Name	**Language(s)**	**# Hours**	**Model Type**	**Link**
1	LB-1K-Base	French	1,096	base	LeBenchmark/wav2vec2-FR-1K-base
2	LB-1K-Large	French	1,096	large	LeBenchmark/wav2vec2-FR-1K-large
3	LB-2.6K-Base	French	2,773	base	LeBenchmark/wav2vec2-FR-2.6K-base
4	LB-3K-Base	French	2,933	base	LeBenchmark/wav2vec2-FR-3K-base
5	LB-3K-Large	French	2,933	large	LeBenchmark/wav2vec2-FR-3K-large
6	LB-7K-Base	French	7,739	base	LeBenchmark/wav2vec2-FR-7K-base
7	LB-7K-Large	French	7,739	large	LeBenchmark/wav2vec2-FR-7K-large
8	F-1K-Base	French	1,041	base	LeBenchmark/wav2vec-FR-1K-Female-base
9	F-1K-Large	French	1,041	large	LeBenchmark/wav2vec-FR-1K-Female-large
10	M-1K-Base	French	1,006	base	LeBenchmark/wav2vec-FR-1K-Male-base

11	M-1K-Large	French	1,006	large	LeBenchmark/wav2vec-FR-1K-Male-large
12	Tamasheq	Tamasheq	243	base	https://github.com/gruly/IWSLT2022_Tamasheq_data
13	Tamani Kalangou	Tamasheq, Hausa, Fulfulde, French, Zarma	641	base	https://github.com/gruly/IWSLT2022_Tamasheq_data
Soon to be Available Models					
14	LB-14K-Large	French	14,000	large	
15	LB-14K-xlarge	French	14,000	large	
16	SELMA	Multi-lingual	6,000	large	

Table 1 *List of Trained wav2vec 2.0 Models*

Models 1 to 7 were trained in the context of the LeBenchmark initiative[1] in which the SELMA project was involved through the LIA partners (see our paper[2] for the details). We trained massive wav2vec2.0 models for the French language using diverse audio data and two architecture sizes. These models are freely available at HuggingFace hub[3], and they will provide us with a base for transfer learning approaches for speech.

Models 8 to 11 were recently trained in our investigation regarding gender bias in SSL models for speech processing. We train models on female and male voice only, and we study how this setting of extreme unbalance of pre-training data impacts the performance on posterior speech-to-text systems. These models will soon be publicly available at Hugging Face.

Models 12 and 13 focus on Nigerian languages, and they were part of the IWSLT 2022 speech translation campaign, low-resource track. With these models, our investigation focuses on understanding if training SSL models on languages geographically close, and with known

[1] http://lebenchmark.com

[2] https://openreview.net/forum?id=cSYVlEL57gK

[3] https://huggingface.co/LeBenchmark

lexical borrowing (model 13), can be a solution for the shortage of data in one given language (model 12). We intend to release these models on a dedicated Hugging Face webpage soon.

Models 14 and 15 focus on French language with the use of more data for the wav2vec2.0.

Model 16 focuses on multilingual dataset collected and provided by the Deutsche Welle partner. The domain of the data is broadcast news.

2.2 Automatic Speech Recognition (ASR)

End-to-end automatic speech recognition (ASR) models have been built in the framework of the SELMA project. Some members of the LIA partner are strongly involved in the development of the SpeechBrain project (https://speechbrain.github.io), and this toolkit is used by LIA to develop its new ASR model for the SELMA framework. These ASR models are mainly built on the use of pretrained wav2vec 2.0 models: some of which are described in the previous section.

For now, these programs are research tools; therefore, integration work is still necessary to make most of them accessible to a non-specialist public. Some of this software has been released on the SELMA GitHub, https://github.com/SELMA-project/LIA_speech/asr.

The ASR models built in the framework of SELMA in 2021 until March 2022, and available in this repository, target the following languages:

1. Brazilian Portuguese
2. French
3. Modern Standard Arabic
4. Tunisian Dialect

During the last months, the LIA partners developed a federated learning (FL) ASR system, which is also based on using the wav2vec2.0 model. FL is a distributed machine learning paradigm that aims to train a machine learning model without data sharing collaboratively. It consists in a network of multiple clients and one server. SpeechBrain and Flower toolkits have been used. Flower[4] is an open-source framework that allows us to build FL experiments and

[4] https://flower.dev

considers the highly varied FL facility scenarios. One member of the LIA partner is involved in the development of this toolkit.

We recently submitted a paper to the ICASSP conference[5]. We intend to release upon paper acceptance this ASR system with a complete recipe (data processing, ASR training, and evaluation scripts).

2.3 Speech Translation (ST)

During this first year of the SELMA project, we focused on assessing the capability of speech translation models in extremely low-resource settings. With this goal, we have used the Tamasheq language as a use case. While this language is not part of the collection of languages initially targeted by the SELMA project, it allows us to assess the state-of-the-art performance in similar settings to many low-resource languages targeted by our project.

We recently submitted our best speech translation model to the IWSLT 2022 Speech Translation Challenge[6], and we now intend to use the lessons learned from this research challenge to develop similar models for languages such as Pashto and Hausa. Our submission was based on the wav2vec 2.0 model 12 in Table 1, and it explored intermediate representations from this SSL model's transformer encoder stack in order to reduce the number of trainable parameters. This way, we attenuated the impact of fully fine-tuning this model in low-resource settings, achieving better results. We intend to release this architecture soon on SpeechBrain, together with a companion paper. Language-dedicated models are left for the following software delivery.

The submitted Tamasheq to French system developed by the LIA partner has been the best system during this challenge compared to the other submitted systems. It is based on the use of wav2vec 2.0. This system has been released by LIA at SpeechBrain's GitHub page[7].

LIA continue to work in Speech Translation and plans to participate in the IWSLT 2023 for the two tasks of IWSLT 2022 and a new task related to the Pashto languageage. We will release the systems trained during the IWSLT 2023.

---

[5] https://2023.ieeeicassp.org

[6] https://iwslt.org/2022/low-resource

[7] https://github.com/speechbrain/speechbrain/tree/develop/recipes/IWSLT22_lowresource

2.4 Named Entity Recognition from Speech (NER-S)

As for the previous tasks, the SpeechBrain toolkit was used to build a system for the MEDIA French corpus. This corpus is a dataset of phone audio recordings with manual annotations dedicated to semantic concepts extraction (SCE) from the speech in the context of human/machine dialogues. The corpus contains manual transcriptions and semantic annotations of dialogues from 250 speakers and totals less than 25 hours of speech. The semantic concepts extraction task is close to the named entity recognition from speech (NER-S) task, both slot-filling tasks. The main difference comes from the semantic annotation, which is more generic for the NER-S task and more specific for the SCE task (a named entity is defined as a snippet of the global information contained in a document, while a semantic concept is defined for a specific task).

A recipe (including data preparation, training, and evaluation scripts) for the MEDIA corpus (ASR and SLU tasks) has been built and tested, which will be later integrated into the SpeechBrain toolkit (https://github.com/speechbrain/speechbrain/tree/develop/recipes). Its integration is not yet finalized due to a minor update needed for the data processing in the future evolution of the MEDIA dataset (see pull request: https://github.com/speechbrain/speechbrain/pull/1172). SpeechBrain will permit the code to be persistent, thanks to community maintenance.

2.5 Text-to-Speech Synthesis (TTS)

During the first semester of the SELMA project, we released the first version of our TTS engine. It was a two-part system composed of an acoustic model and a vocoder. The acoustic model generates acoustic features from linguistic features (text in our case), and the vocoder synthesizes waveform from the acoustic features. For the acoustic model, we used Tacotron 2 with WaveRNN vocoder.

During the second semester, we mainly worked on improving our baseline system in terms of robustness and inference time. To do this, we considered moving from our two-part system to an end-to-end model. This has the advantage of reducing error propagation due to the cascading system. On the other hand, using an end-to-end model gives us a faster inference time, which is very important since the model is deployed in production.

We found that variational autoencoder-based topology matches perfectly with our requirements. We have conducted several experiments that have shown that we can replicate

the performance of Tacontron 2 + WaveRNN while decreasing the inference time by at least 150 times. The TTS API is accessible through the plainX platform and the docker image can be downloaded from our cloud page.

To train the speech synthesis engine, we use the audio news bulletins that are produced by DW's Brazil department. The audio files have been downloaded from YouTube and the scripts were retrieved from GitHub in a repository with all the text scripts that DW uses to produce their weekday news podcasts. The dataset contains approximately 32 hours of speech from 8 speakers.

2.6 Punctuation and Capitalization Recovery (PCR)

The majority of conversational systems are not able to produce formatted and punctuated transcripts, making the text difficult to read and understand. This lack of formatting can hinder comprehension, even if the transcription is free of errors. Our target is to use a multi-task system that can leverage the connection between punctuation and capitalization to improve the performance of transcriptions.

In this release, we propose a multi-task approach for joint true casing and punctuation prediction using pre-trained BERT[8] models. Previous research has demonstrated the effectiveness of fine-tuning pre-trained models for NLP tasks, so we follow a similar method for our true casing and punctuation prediction tasks. The BERT base model consists of a sequence of self-attention and fully connected layers and is trained on the masked language model and next sentence prediction objectives. For fine-tuning, we remove these heads and only use the encoder for further modeling. To make the most of the relationships between true casing and punctuation, we optimize both loss functions together using a multi-tasking approach.

The figure below illustrates the flow chart for our multi-task model using an example. As shown, we process the BERT encoder output using task-specific layers to generate predictions for our tasks. These layers include a dropout layer and a fully connected layer. The BERT encoder serves as a common component for both tasks, encoding information related to both tasks, and then the task-specific layers retain the relevant information for each task. Through optimization of the loss function for each task, we adapt the pre-trained weights and learn new

[8] Devlin, Jacob, Ming-Wei Chang, and Kenton Lee. " BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of NAACL-HLT*. 2018.

parameters. Loss for each task is calculated by comparing the output of the fully connected layers against the corresponding ground truth labels. We use the cross-entropy loss function to compute losses, as the targets for both tasks are categorical.

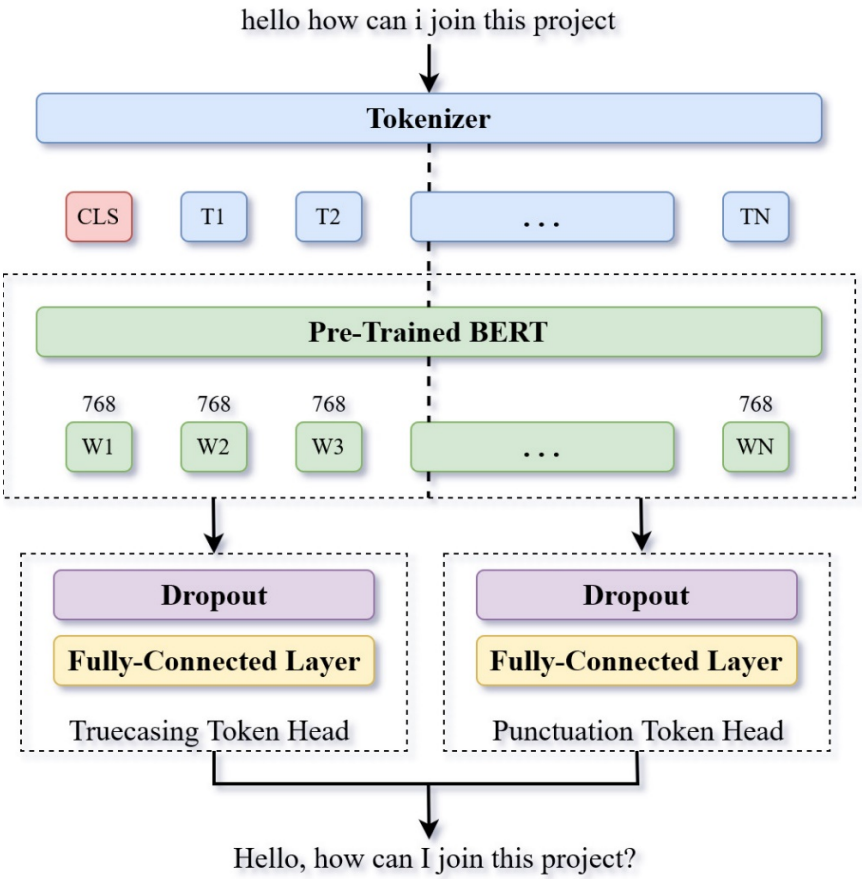

Figure 1 *Block Diagram of Joint Punctuation and True Casing Modeling*

In our experimental evaluations, we used the Tatoeba[9] dataset, which is a collection of translated sentences and is commonly used as a benchmark especially for machine translation. This dataset includes text data for thousands of language pairs across over 500 languages. For each word, we assigned a true casing and punctuation label. For true casing, we considered two labels: lower and upper casing. For punctuation, we assigned one of four possible labels for each word: blank, full-stop, comma, or question mark. In cases where punctuation appeared after a word, we assigned that word with that punctuation, otherwise we assigned it a blank label.

[9] Tatoeba Project: https://tatoeba.org/en

We evaluated all our models using macro F1-scores, including both cased and uncased variations of BERT. In general, we found that the uncased model performed better than the cased variant for this task, which was expected as we lowercased the entire corpus to reduce bias in period prediction. For all our models over German, English, Spanish, French, Italian, Latvian and Portuguese languages, we selected the best-performing ones on the validation set based on loss and macro F1-score, then evaluated them independently on the test set.

3. Future Work

In future releases, we will first investigate language scaling to achieve decent performances over more languages. Especially multi-lingual models that can handle several linguistic contents into a single system will be useful for the whole SELMA languages. For PCR, the current model only supports the most used punctuations. This choice has been made due to the lack of currently available data, but more punctuation marks will be added to the system in the future through more diverse data sets. Apart from that, an audio-driven approach will also be investigated. Since speech prosody reflects the information structure of the speech to some extent, features representing intonation, stress, pitch, energy, and pausing could be utilized together with punctuation modeling.