Research and Innovation Action (RIA) H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu

D3.2 Initial release of transcription, punctuation, translation, voice synthesis capabilities

Work Package	3
Responsible Partner	LIA
Author	Yannick Estève
Contributors	Antoine Caubrière, Jarod Duret, Tilo Himmelsbach, Gaëlle Laperrière, Salima Mdhaffar, Tugtekin Turan, Marcely Zanon Boito
Reviewer	Tugtekin Turan
Version	V1.0
Contractual Date	31 March 2022
Delivery Date	31 March 2022 (Resubmitted 25 May 2022)
Dissemination Level	Public

Version History

Version	Date	Description
0.1	15/03/2022	Initial Table of Contents (ToC)
0.2	29/03/2022	Main input from partners
0.3	30/03/2022	Merging
0.4	31/03/2022	Internal Review version
0.5	31/03/2022	Finalization
1.0	31/03/2022	Publishable version
1.1	17/05/2022	Unmerged deliverables
2.0	25/05/2022	Publishable version

Executive Summary

This initial deliverable describes the first release of software components developed within WP3 and the integration with the SELMA orchestration platform at this development stage of very novel research tools.

SELMA's approach to speech and language processing is targeting both low and high resourced languages

This document also provides an update of post-editing and user feedback capabilities to be followed by the interim and final releases later in the project.

Table of Contents

1. Introduction

In this report, we detail the models and software built by the SELMA partners and released for internal or external purposes, depending on the maturity of the software. Most of them are very novel and, at this stage, are still in the state of research tools. These tools have proven their worth in experimental setup and have advanced the state-of-the-art in terms of accuracy. To move towards the deployment of these tools, it is now necessary to work on their integration into the targeted platforms. The software addresses:

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

We also present in this document our first software that allows us to inject linguistic information from text into an end-to-end neural ASR model. To our knowledge, this is the first software that offers this possibility for such technology. All components are deployed as containers and will be available at our Docker hub, https://hub.docker.com/orgs/selmaproject.

2. Released Software

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
Soon to be Available Models					
8	F-1K-Base	French	1,041	base	
9	F-1K-Large	French	1,041	large	
10	M-1K-Base	French	1,006	base	
11	M-1K-Large	French	1,006	large	
12	Tamasheq	Tamasheq	243	base	
13	Tamani Kalangou	Tamasheq, Hausa, Fulfulde, French, Zarma	641	base	

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 Hugging Face 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.

Lastly, 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 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.

2.2 Automatic Speech Recognition

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.

[1] http://lebenchmark.com

[2] https://openreview.net/pdf?id=TSvj5dmuSd

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

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

- Brazilian Portuguese
- French
- Modern Standard Arabic
- Tunisian Dialect

2.3 Punctuation and Capitalization Model

Ensuring proper punctuation and letter casing is a critical post-processing step toward applying machine translation or automatic speech recognition. In this initial version, we present a transformer-based automatic punctuation and capitalization model that accepts lexical information (the words themselves) and outputs a text with improved readability. The implemented method consists of pre-trained Bidirectional Encoder Representations from Transformers (BERT) followed by two token classification heads. One classification head is responsible for the punctuation task, the other one handles the capitalization task. Such architecture allows this model to solve two tasks at once with only a single pass through the BERT.

Models were trained using NVIDIA's NeMo toolkit, which has an Apache-2.0 license. The program is available in the SELMA GitHub: https://github.com/SELMA-project/punctuation-capitalization-recovery

2.4 Speech Translation

During this first year of the SELMA project, we focused on assessing the capability of speech translation models in extreme low-resource settings. With this goal, we have been using the Tamasheq language as a use case. We highlight that, 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[4], 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.

2.5 Named Entity Recognition (and Semantic Concept Extraction) from Speech

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 speech in a 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. Semantic concepts extraction task is really close to the named entity recognition from speech (NER-S) task, being 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 to 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 coming 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.

The recipe's results are close to the end-to-end system's state-of-the-art ones. This proves the recipe is operational and simply needs tuning to enhance the results.

[4] https://iwslt.org/2022

2.6 Text-to-Speech Synthesis

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[5] with WaveRNN[6] 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, by using an end-to-end model, we have a faster inference time, which is very important since the model is deployed in production.

We found that variational autoencoder[7] 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[8].

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.

[5] https://arxiv.org/abs/1712.05884

[6] https://arxiv.org/abs/1802.08435

[7] https://github.com/jaywalnut310/vits

[8] https://drive.google.com/drive/folders/14g4qbM_F4Mn5m8bVxiUCLEWLasll2aiW

3. Future Plan

The programs released this year by the SELMA project are still research (and so very recent) tools with very nice results in terms of accuracy, covering the SELMA WP3 tasks. Some of them, like the TTS software, were mature enough to be integrated into the SELMA platforms.

In the next months, the other programs will be packaged and profiled to be integrated in these platforms. At the same time, efforts will be made to extend the language coverage to match the SELMA objectives.