21 - LiSTra Automatic Speech Translation: English to Lingala case study

Salomon Kabongo,
Vukosi Marivate
Herman Kamper

Abstract

In recent years there have been great interests in addressing the low resourcefulness of African languages and providing baseline models for different Natural Language Processing tasks (Orife et al., 2020). Several initiatives (Nekoto et al., 2020) on the continent use the Bible as a data source to provide proof of concept for some NLP tasks. In this work, wepresent the Lingala Speech Translation (LiSTra) dataset, release a full pipeline for the construction of such dataset in other languages, and report baselines using both the traditional cascade approach (Automatic Speech Recognition -> Machine Translation) and a revolutionary transformer-based End-2-End architecture (Liu et al., 2020) with customized interactive attention that allows information sharing between the recognition decoder and the translation decoder.

1. Introduction

Automatic Speech Translation (AST) is the task of converting an utterance from a source language to transcription in a target language. Success in this task will revolutionize online education among other things as the majority of educational content available on e-learning platforms are mainly English-centric which is a bottleneck to people with limited or no knowledge of English. Inspired by (Orife et al., 2020) we performed an AST proof-of-concept from English to Lingala.

One bottleneck in experimenting with ASR on low resources languages has been lack of aligned data, we introduce the **Ligala Speech Translation (LiSTra)** a dataset of reading of the Bible, its transcription, and the Lingala translation. The choice of the bible as a data source is motivated by missionary work on the African continent, which made available the transcription and translation alignments.

The traditional approach in AST consist of first doing Automatic Speech Recognition (ASR), then feeding the output into a Machine Translation (MT) system, one pitfall in this approach is the error propagation (not back-propagation) that arise due to the fact that the 2 components are trained independently. In this work, we will release a baseline for AST both in a pipeline (ASR -> MT) as well as in an end-to-end setting.

2. Dataset

Let $\mathbf{D} = \{\mathbf{S}^j, \mathbf{E}^j, \mathbf{L}^j\}_{j=1}^{|D|}$ the dataset that we would like to create, with **S** *the speech utterance*, **E** the corresponding *transcription* and **L** the *lingala translation*.

LiSTra is a systemic crawl of the Bible both at from the jw.org for Lingala translation and bible.is for speech and English transcription. WebMAUSBasic service of the Bavarian Archive for Speech Signals (BAS) was used to split the audio from the chapter level to verses level.

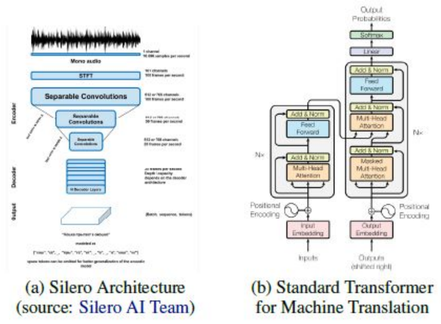
4. Experiments

		LiSTra		
Text language Source	Split	Examples	Avg. text length	Total Unique Words
English (En)	train	23717	24.2712	13139
	test	5930	24.2076	7772
Text language Target	Split	Examples	Avg. text length	Total Unique Words
Lingala (ln)	train	23717	25.9165	16808
	test	5930	25.7489	8940
Speech Source	Split	Examples	Avg. audio length (seconds)	Total numb. hours
English (.wav)	train	23717	9.2880	61
	test	5930	9.2715	15

Table 1: Data statistics of LiSTra

4.1. AST: Cascade

The Cascade architecture is made of two separate models as described in figure 1, a pre-trained Sirelo11 Model and a traditional transformer-based Machine translation architecture which receive the output of the former one to perform Speech Translation.



(a) Silero Architecture (source: Silero AI Team)

(b) Standard Transformer for Machine Translation

Figure 1: Cascade Approach : Speech Recognition (a) + Machine Translation (b)

4.2. AST: end-2-end

In the end-2-end setting, we use a transformer based model, that is made of one encoder and two decoders as shown in the figure 2. This architecture has shown promising results recently (Liu et al., 2020) specially due to the interaction between the recognition decoder and the translation decoder.

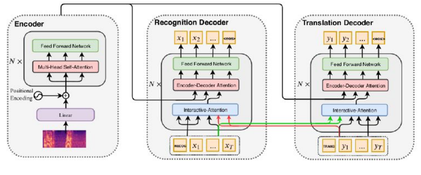


Figure 2: Synchronous AST Architecture (Liu et al., 2020)

5. Results and Conclusion

The Cascade architecture is made of two separate models, a pre-trained Sirelo Model and a traditional transformer-based Machine translation architecture which receive the output of the former one to perform Automatic Speech Translation.

We pre-trained the Machine Translation model on the JW300 dataset (Agic and Vulic, 2020) and train further on LiSTra data. The recognized waves from silero are then fed into the trained MT to obtain our Speech translation output.

In the end-2-end setting, we used a transformer-based model, that is made of one encoder and two decoders. This architecture has shown promising results recently (Liu et al., 2020) especially due to the interaction between the recognition decoder and the translation decoder.

Given that the Text to Speech task is often more difficult than Automatic Speech Recognition, we used (Liu et al., 2020) the *wait-k* policy approach that basically allows waiting for a certain time to allow the recognition decoder to transcribe some words before it can start translating.

	wait-1			wait-2			wait-3		
Architecture	WER ↓	BLEU (en) ↑	BLEU (ln) ↑	WER ↓	BLEU (en) ↑	BLEU (ln) ↑	WER ↓	BLEU (en) ↑	BLEU (ln) ↑
Pipeline[13]	8.27	84.90	13.92	x	x	x	x	x	x
End-2-End	8.06	84.40	26.45	**7.81**	**84.90**	**28.52**	7.87	84.73	26.99

Table 2: Results : Experimentation for different value of k

In this work, we presented LiSTra, the first dataset for automatic speech translation from English to Lingala, and a pipeline to allow researchers working on low-resource languages to create a similar dataset for their language. *Despite the dataset being biased toward religious content this can serve as a starting dataset for proof-of-concept*. In addition, we reported baselines in both Pipeline and End-2-End architecture and concluded that the End-2-End architecture performs quite well despite the limited amount of data.

References

Željko Agic and Ivan Vulic. 2020. Jw300: A wide coverage parallel corpus for low-resource languages.

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020. Synchronous speech recognition and speech-to-text translation with interactive decoding. In AAAI, pages 8417–8424.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus,Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane machine translation for africa. arXiv preprint arXiv:2003.11529.

Wilhelmina Nekoto, Vukosi Marivate,et al. 2020 Participatory research for low-resourced machine translation: A case study in African languages. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2144–2160.