Using ASR-Generated Text for Spoken Language Modeling

Nicolas Hervé[1] Valentin Pelloin[2] Benoit Favre[3] Franck Dary[3]
Antoine Laurent[2] Sylvain Meignier[2] Laurent Besacier[4]

[1]Institut National de l'Audiovisuel (INA), France [2]Laboratoire d'Informatique de l'Université du Mans (LIUM), France
[3]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France [4]Naver Labs Europe (NLE), Meylan, France

Introduction

- large Language Models (LM) are trained with texts which do not reflect well spoken language
- large LM are used for spoken language downstream tasks
- the French National Audiovisual Institute (INA) have a collection of 350,000 hours of TV Shows
- transcribe and use these 350,000 hours to train or finetune a large LM
- evaluate on word prediction and speech downstream tasks

From FlauBERT to FlauBERT-Oral

- transcription of $350,000$ hours of spoken language
  - Kaldi-based ASR system
  - vocaburary of 160k most frequent words
  - trained on 440 hours of transcribed speech
  - 350,000 hours of French TV and Radio Shows ≈ 19GB of ASR-generated text
- FlauBERT baseline (**FBU**): flaubert-base-uncased [1]
- FlauBERT-O-base_uncased (**FT**): fine-tuning of FBU model using our ASR text
- FlauBERT-O-mixed (**MIX**): re-trained LM using ASR text + written text
- FlauBERT-O-asr: re-trained LM using ASR text only
  - **ORAL**: with initial BPE tokenizer provided with *flaubert-base-uncased*
  - **ORAL_NB**: with new BPE tokenizer obtained from ASR text

Word Prediction Task

Experiments on external written and oral corpuses :

- **afp2021** : AFP dispatches from the year 2021
- **parl_13** and **parl_15** : manual transcripts of the French National Assembly sessions (under Sarkoy and Macron)
- **studio_manual** and **studio_asr** : transcripts of an Ina corpus of both educational and interviews videos. We have very good manual transcripts as well as the transcripts from our ASR system

Corpus	FBU	FT	MIX	ORAL_NB	ORAL
afp2021	53.1	55.1	**60.9**	48.6	51.9
parl_13	**64.9**	63.6	64.5	58.8	60.0
parl_15	**64.6**	64.3	64.3	59.7	60.7
studio_asr	40.2	**48.0**	46.6	46.8	47.2
studio_manual	57.0	**59.4**	58.9	56.3	56.9

Table 1. Word prediction task accuracies

Automatic Classification of TV Shows

- Corpus : 47 867 short TV shows, average length 92 seconds, from 4 TV channels (TF1, France 2, France 3 and M6), annotated by INA's documentalists into 14 categories
- Classification experiments : baseline (SVM and TF-IDF representations) and our oral models (using the HuggingFace Transformers library and a simple dense classification layer on top of our models)

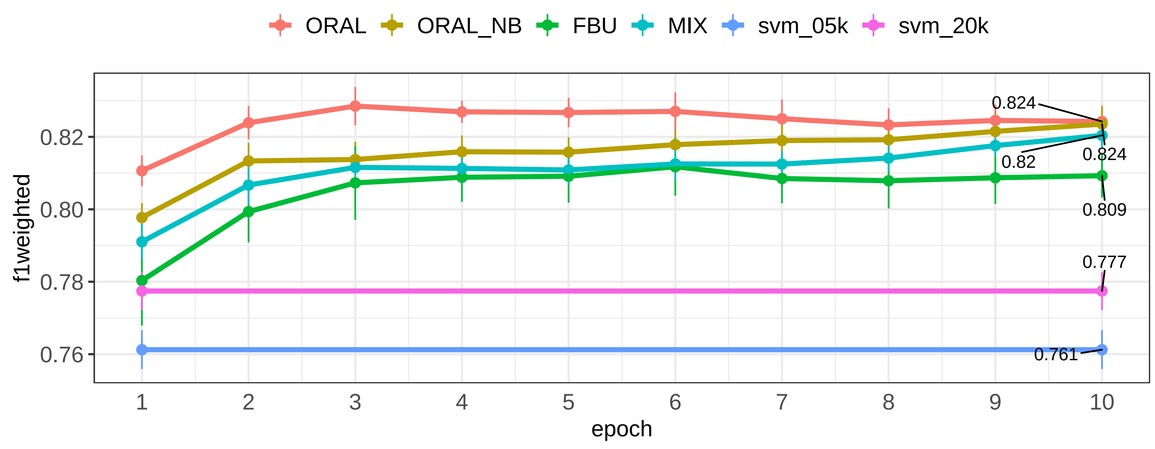

Figure 1. TV news classification - **Standard setting**

Syntactic Analysis of Spoken Conversations

Transition Based Parser jointly preditcs POS tags and dependency tree. Taking as input word embeddings either: • Not pretrained • From Fasttext • From one of our 4 competing models. Higher is better, bold is best:

Repr.	LAS	UAS	UPOS
No pretraining	84.92	88.48	94.51
Fasttext	85.36	88.76	95.12
FBU	85.55	89.02	93.36
MIX	86.33	89.79	94.43
ORAL	**87.65**	**90.92**	95.55
ORAL_NB	87.54	90.73	**95.63**

Conclusion and future work

- new models for French (FlauBERT-O) are shared
- spoken language modeling using ASR generated text is possible
- FlauBERT-O is generally better than initial FlauBERT for downstream speech tasks
- more downstream speech tasks should be evaluated

Reference

[1] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May 2020. European Language Resources Association.