Speech @ BigScience

Analyse syntaxique de la parole

Benoit Favre & Franck Dary - JTT 2022-01-20

Contexte : BigScience

https://bigscience.huggingface.co

Projet de 1 an avec 450 chercheurs de 50 pays (porté par Huggingface)

Objectif : faire un GPT-3 multilingue, sans biais, open-source

Workshops tous les 3 mois

- Prochain à ACL 2022, 27 mai : “Workshop on Challenges & Perspectives in Creating Large Language Models”

Jusqu’ici : Modèle T0 (16x plus petit que GPT-3) avec du fine tuning sur des tâches diverses pour tester la généralisation à de nouvelles tâches (https://arxiv.org/abs/2110.08207)

Modèles de langage pour la parole

Collaboration LIS (B. Favre et F. Dary), NaverLabs (L. Besacier) INA (N. Hervé), LIUM (A. Laurent, V. Pelloin), Orange (G. Lecorvé et G. Damnati)

- À côté de quoi passe-t-on quand on entraîne un modèle de langage sur une tranche d'internet ?
- Le fait que la parole soit sous-représentée est-il un problème ?
- Les particularités de l'oral sont-elles capturées par les modèles ?
- Peut-on utiliser des transcriptions automatiques pour apprendre un ML ?
- Peut-on faire un modèle qui soit bon à la fois sur du texte et de la parole ?
- Comment combiner les modèles textes et acoustiques (i.e. HuBERT) ?

Méthodologie

Source de transcriptions : Émissions de télévision transcrites automatiquement en grande quantité

Modèles de langage

- Baseline : FlauBERT pour le français
- Mixed : FlauBERT entraîné sur un mix de transcriptions auto et de textes
- ASR : FlauBERT entraîné sur des transcriptions auto

Évaluation sur des tâches aval (syntaxe, classification thématique, SLU, entités nommées, reranking d’ASR…)

FlauBERT (Hang Le et al, LREC 2020)

Même architecture que BERT (base = 12 couches, 12 têtes, 137M paramètres)

Entraîné sur 24 sous-corpus (71G de texte, ~50% Common Crawl), 50k BPE

Évaluaiton FLUE

Tâche	Classification			Paraphrase	NLI	Constituants		Dépendances		Désambiguïsation	
Section	Livres	DVD	Musique							Noms	Verbes
Mesure	Acc.	Acc.	Acc.	Acc.	Acc.	F_1	POS	UAS	LAS	F_1	F_1
État de l'art ant.	91.25[c]	89.55[c]	93.40[c]	66.2[d]	80.1/**85.2**[e]	87.4[a]		89.19[b]	85.86[b]	-	43.0[h]
Sans pré-entr.	-	-	-			83.9	97.5	88.92	85.11	50.0	-
FastText	-	-	-			83.6	97.7	86.32	82.04	49.4	34.9
mBERT	86.15[c]	86.9[c]	86.65[c]	89.3[d]	76.9[f]	87.5	98.1	89.5	85.86	56.5	44.9
CamemBERT	**93.40**	**92.70**	94.15	89.8	81.2	88.4	**98.2**	91.37	88.13	56.1	**51.1**
FlauBERT$_{BASE}$	**93.40**	92.50	**94.30**	**89.9**	81.3	**89.1**	98.1	**91.56**	**88.35**	54.9/**57.9**[g]	47.4

Sources de données audio

Sources

- Émissions TV collectées à l'INA (16 chaines, 350k heures, 2018-2020)
- ~13G de transcriptions automatiques

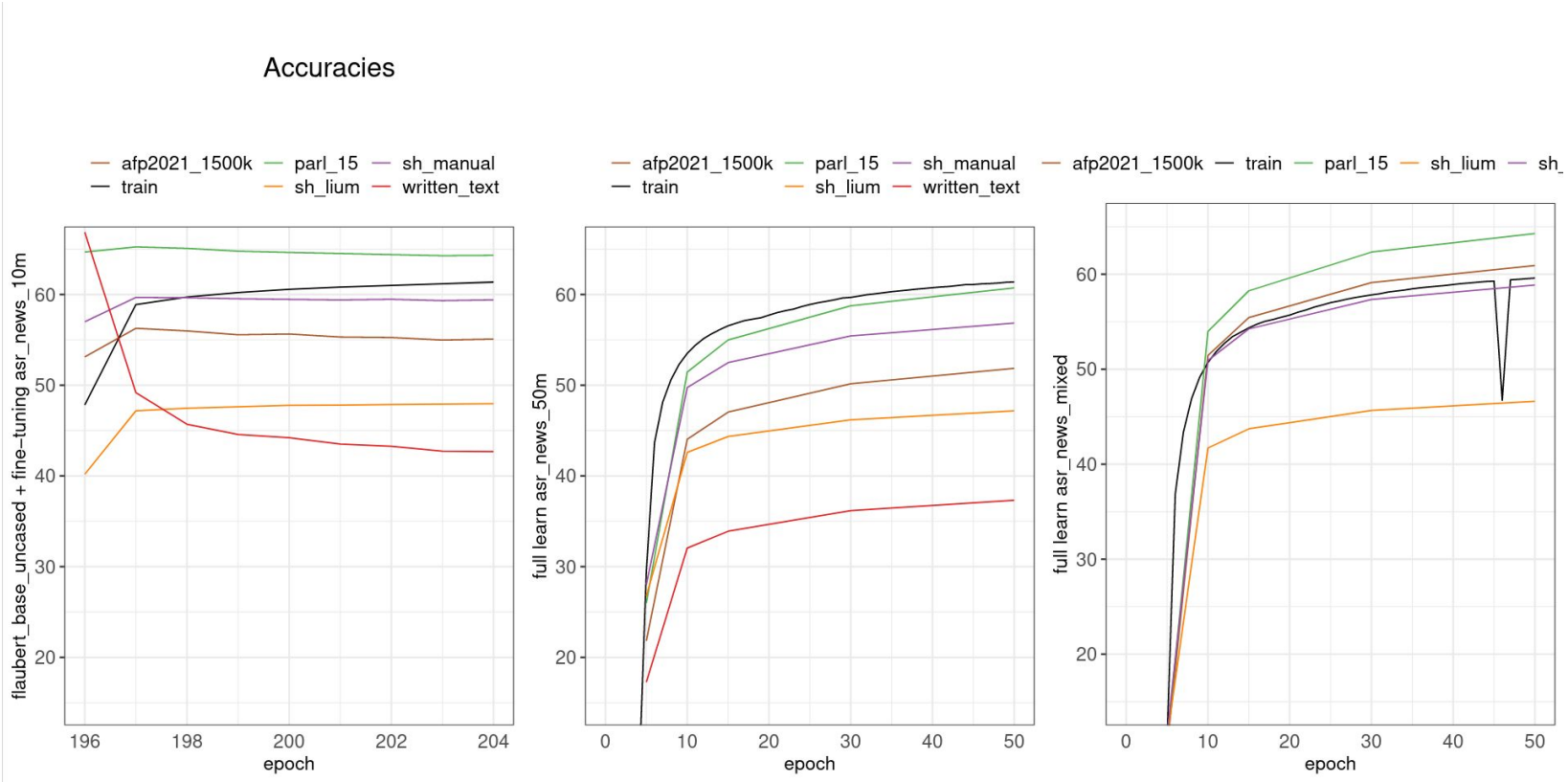
Transcription

- Système de transcription du LIUM : Kaldi (HMM-TDNN, MMI) vocabulaire de 150k mots (Povey et al, 2016)

Entraînement FlauBERT

- Par tour de parole, tronqué à 512 tokens,
- Textes en minuscules et sans ponctuation

Précision du modèle de langage (tâche MLM ?)



Tâche aval : analyse syntaxique

Corpus Orféo

- 370h d'audio au total (3M mots)
- Annotation en syntaxe : ~25h, 150k tokens pour l'entraînement, 20k dev, 30k test
- Segmentation et transcription de référence
- Mots en minuscules, pas de ponctuation
- Traitement particulier des noms propres anonymisés

Analyseur syntaxique (Dary et al, 2021)

- Système incrémental Macaon, un LSTM+MLP décide des actions shift/réduce
- Les représentations sont calculées pour un contexte donné (max 512 tokens) et injectées *in situ* des embeddings de mots
- Elles ne sont **pas affinées** lors de l'entraînement
- Les tokens divisés par BPE sont moyennés

Résultats bruts

Baselines :

- Apprentissage des embeddings initialisés aléatoirement (No pretrain)
- FastText-300 (OOVs inférés à partir des n-grammes de caractères en test)

System	LAS	UAS	UPOS
No pretrain	84.92	88.48	94.51
FastText-300	85.36	88.76	95.12
FlauBERT Baseline	85.55	89.02	93.36
FlauBERT Mixed	86.33	89.79	94.43
FlauBERT ASR	**87.65**	**90.92**	**95.55**

Qu’est-ce qui fait la différence ?

Impact du contexte et de la ponctuation

Choice of context

	LAS	UAS	UPOS
FlauBERT Baseline (dialog)	**85.55**	**89.02**	**93.36**
FlauBERT Baseline (turn)	74.17	80.20	80.60
FlauBERT Baseline (sentence)	70.63	76.65	77.97

Punctuation

	LAS	UAS	UPOS
FlauBERT Baseline	85.55	89.02	93.36
FlauBERT Baseline punc	87.48	90.69	95.03
FlauBERT ASR	**87.65**	**90.92**	**95.55**

- Les représentations sont fortement dégradées si l'on utilise pas le contexte le plus large
- L'absence de ponctuation semble être un facteur central

Impact des OOV de l'ASR ?

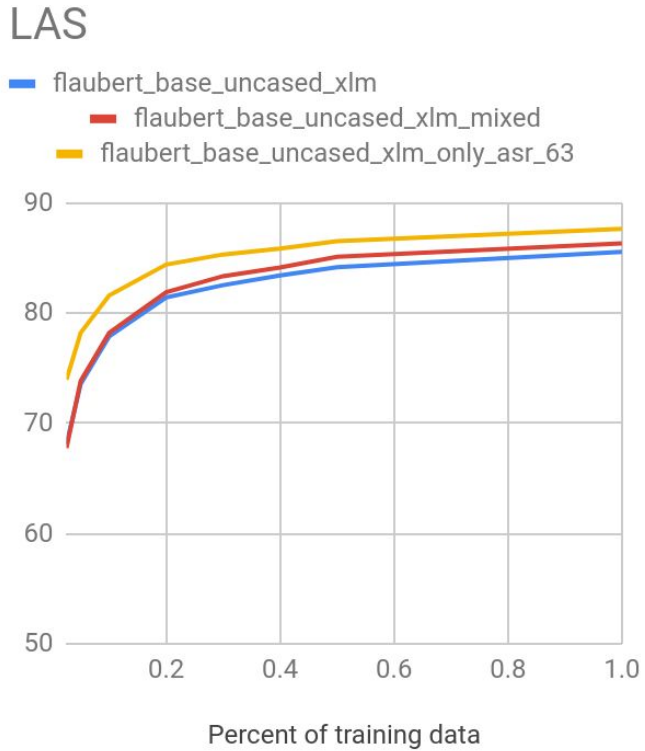
	LAS			UAS			UPOS		
	Global	OOV	Diff	Global	OOV	Diff	Global	OOV	Diff
Baseline	85.55	74.10	-11.45	89.02	82.20	-6.82	93.36	79.00	-14.36
Mixed	86.33	74.40	-11.93	89.79	82.47	-7.33	94.43	80.35	-14.07
ASR	**87.65**	73.68	-13.97	**90.92**	82.81	-8.11	**95.03**	81.11	-13.92

ASR : lexique de taille 150k; Flaubert : BPE avec 50k tokens

- Perte plus grande sur les OOV pour le modèle ASR
- Perte moins forte en UAS que LAS

Évolution lorsqu'on a peu de données d'apprentissage ?

- Les représentations ASR semblent meilleures, en particulier lorsqu'il y a peu de données d'apprentissage.
- Le modèle mixed n'est pas vraiment bon.
- Courbe similaire pour UAS et UPOS



Perspectives

En cours

- Refaire les expériences sur l'ASR des données Orfeo
- Se comparer avec les résultats sur les autres tâches

À plus long terme

- Évaluer sur des phénomènes propres à l'oral comme les disfluences ou les marqueurs de discours
- Tester d'autres ASR moins dépendants du lexique
- Intégrer toutes ces idées dans les gros modèles Bigscience