Reunion 24/10

To do list

- Vectoriser les POS / chunks
- Investiguer la prediction des pauses à l'oral
- Distribution random des pauses
- Equilibrer les données:
  - utiliser autre corpus que celui dans planification
  - essayer d'autres ratios pour l'augmentation des données

1. Vectoriser les POS

Pour integrer la vectorisation POS dans le code:

- Avant de tokeniser avec CamemBERT, j’ai utilisé Spacy pour generer les POS
- J’ai vectorisé les POS
- J’ai *essayé* de concantener les POS embeddings.

Problème d'alignement

La tokenisation entre CamemBERT et Spacy ne s'aligne pas toujours de manière cohérente, entraînant des difficultés lors de la concaténation des embeddings.

Plutôt que de concaténer les embeddings POS et les tokens de CamemBERT, j'ai choisi une approche plus simple consistant à remplacer directement les mots par leurs POS correspondants. Cela permet de contourner le problème d'alignement tout en intégrant l'information grammaticale dans le modèle.

Example d’output:

Original Text:
Cependant, cat_4 pour les personnes ne voyant tout d'abord aucun avantage personnel cat_4 , ne trouve cat_1 nt aucune satisfaction cat_5 dans celle-ci. Ainsi, cela peut entrainer un sentiment de rejet et d'incompréhension. cat_5

Text with POS Tags:
ADV PUNCT PROPN SPACE ADP DET NOUN ADV VERB ADV ADP ADV DET NOUN ADJ ADJ PUNCT ADV VERB NOUN CCONJ DET NOUN SPACE VERB ADP PROPN PUNCT ADV PUNCT PRON VERB VERB DET NOUN ADP NOUN CCONJ ADP NOUN PUNCT ADJ

Results

Les resultats varient à chaque tour.
Le meilleur set:

Accuracy for 'pause': 0.90
Accuracy for 'no pause': 0.92

Mais on a aussi:

	precision	recall	f1-score	support
no pause	0.9761	0.9443	0.9600	1473
pause	0.6150	0.7939	0.6931	165
accuracy			0.9292	1638
macro avg	0.7956	0.8691	0.8265	1638
weighted avg	0.9398	0.9292	0.9331	1638

Par contre, probleme avec la tokenisation des POS par CamemBERT

```
Predictions:
_AD (ADV): pause
V (PUNCT): no pause
_P (PROPN): no pause
UN (SPACE): no pause
CT (ADP): no pause
_PRO (DET): no pause
PN (NOUN): no pause
_SPA (ADV): no pause
CE (VERB): pause
_A (ADV): pause
DP (ADP): no pause
_DE (ADV): no pause
T (DET): no pause
_NO (NOUN): no pause
UN (ADJ): no pause
_AD (ADJ): no pause
V (PUNCT): no pause
_VER (ADV): no pause
```

Pauses randomisées

- J’ai écrit un script pour extraire les pauses et les redistribuer de maniere aleatoire
- J’ai lancé le meme modele qui avait ete entrainé sur les pauses réeles

Randomized test dataset size: 33

Evaluating on Randomized Texts:
Accuracy for 'pause' on randomized texts: 0.31
Accuracy for 'no pause' on randomized texts: 0.72

Randomized Test Accuracy:
{'pause_accuracy': 0.3098236775818639 7,
'no_pause_accuracy': 0.7201411509229099}

Resultat

La performance du modèle sur des textes aléatoires montre une baisse significative dans la détection des pauses (31 % de précision).

Cela suggère que le modèle est plus sensible à la structure originale des pauses qu'à une distribution aléatoire.

Prédiction des pauses dans les corpus oraux :

1. Détection acoustique :

Silence Duration: Pauses sont souvent identifiées par la durée des silences. Typiquement, les silences de plus de 200 ms sont considérés comme des pauses.

Prosodic Cues: Les variations dans la prosodie (intonation, rythme) telles que des changements de tonalité ou des ralentissements dans la parole peuvent indiquer une pause.

Spectral Analysis: Utilisation de caractéristiques spectrales comme la baisse de l'énergie dans certaines fréquences pour détecter les pauses silencieuses ou non.

2. Approche linguistique :

Étiquetage manuel des pauses: Les transcriptions orales peuvent être annotées manuellement par des linguistes pour identifier les pauses pertinentes, souvent basées sur des critères syntaxiques ou prosodiques.

Indicateurs syntaxiques : Certains mots ou structures syntaxiques (conjonctions, ponctuation orale) peuvent signaler l'apparition de pauses.

3. Utilisation des modèles statistiques :

Modèles HMM (Hidden Markov Models): Utilisés pour segmenter automatiquement la parole et identifier les pauses en se basant sur des transitions probables entre états de parole et de silence.

Modèles à base d'apprentissage profond :

- Types de modèles : Réseaux de neurones récurrents (RNN), réseaux neuronaux à convolution (CNN), ou Transformers.
- Données d’entraînement : Utilisation de grands corpus oraux avec des pauses annotées pour entraîner les modèles à prédire des pauses basées sur des patterns contextuels et prosodiques.

4. Alignement texte-parole :

Les systèmes comme Forced Alignment alignent le texte transcrit à l'audio pour prédire les moments où les pauses devraient se produire en fonction de la durée de silence dans le signal.

Plus specifiquement

Les scores de precision que j'ai vu dans les papiers que j'ai lu etaient assez faibles:

Table 2. Performance (in terms of F-1 score, P: Precision, R: Recall) of various Systems (BS, POS, U) for predictin Pause and Non-pause.

Systems	Pause			Non-Pause		
	R	P	F1	R	P	F1
BS	0.66	0.48	0.58	0.83	0.91	0.86
POS	0.69	0.81	0.75	0.94	0.89	0.91
U	0.70	0.61	0.65	0.91	0.87	0.88
U+SS	0.72	0.68	0.71	0.88	0.89	0.88
POS+SS	0.69	0.79	0.74	0.95	0.91	0.93

https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050915X00202/

https://arxiv.org/pdf/2302.13652

Table 3: Results of position prediction of CPI.

	Precision	Recall	F_β
RPs	0.575	0.261	$F_{0.5} = 0.463$
PIPs	0.848	0.996	$F_2 = 0.962$

Accuracy for 'pause': 0.79
Accuracy for 'no pause': 0.94
Original Text:
Cependant, cat_4 pour les personnes ne voyant tout d'abord aucun avantage personnel cat_4 , ne trouve cat_1 nt aucune satisfaction cat_5 dans celle-ci. Ainsi, cela peut entrainer un sentiment de rejet et d'incompréhension. cat_5

Text with POS Tags:
ADV PUNCT PROPN SPACE ADP DET NOUN ADV VERB ADV ADP ADV DET NOUN ADJ ADJ PUNCT ADV VERB NOUN CCONJ DET NOUN SPACE VERB ADP PROPN PUNCT ADV PUNCT PRON VERB VERB DET NOUN ADP NOUN CCONJ ADP NOUN PUNCT ADJ

Predictions:
_AD (ADV): pause
V (PUNCT): no pause
_P (PROPN): no pause
UN (SPACE): no pause
CT (ADP): no pause
_PRO (DET): no pause
PN (NOUN): no pause
_SPA (ADV): no pause
CE (VERB): pause
_A (ADV): pause
DP (ADP): no pause
_DE (ADV): no pause
T (DET): no pause
_NO (NOUN): no pause
UN (ADJ): no pause
_AD (ADJ): no pause
V (PUNCT): no pause
_VER (ADV): no pause
B (VERB): no pause
_AD (NOUN): no pause
V (CCONJ): no pause
_A (DET): no pause
DP (NOUN): no pause
_AD (SPACE): no pause
V (VERB): no pause
_DE (ADP): no pause
T (PROPN): no pause
_NO (PUNCT): no pause
UN (ADV): no pause
_A (PUNCT): no pause
DJ (PRON): no pause
_A (VERB): no pause
DJ (VERB): no pause
_P (DET): no pause
UN (NOUN): no pause
CT (ADP): no pause
_AD (NOUN): no pause
V (CCONJ): no pause
_VER (ADP): no pause
B (NOUN): no pause
_NO (PUNCT): no pause
UN (ADJ): no pause
              precision   recall f1-score   support

    no pause   0.9761   0.9443   0.9600   1473
       pause   0.6150   0.7939   0.6931   165

    accuracy   0.9292   1638
   macro avg   0.7956   0.8691   0.8265   1638
weighted avg   0.9398   0.9292   0.9331   1638

https://arxiv.org/pdf/2402.13446