Projet PSTALN

Prédiction des *morphy*

Cléa Han, Yanis Labeyrie et Adrien Zabban

15 janvier 2024

Le but : prédire les *morphy*

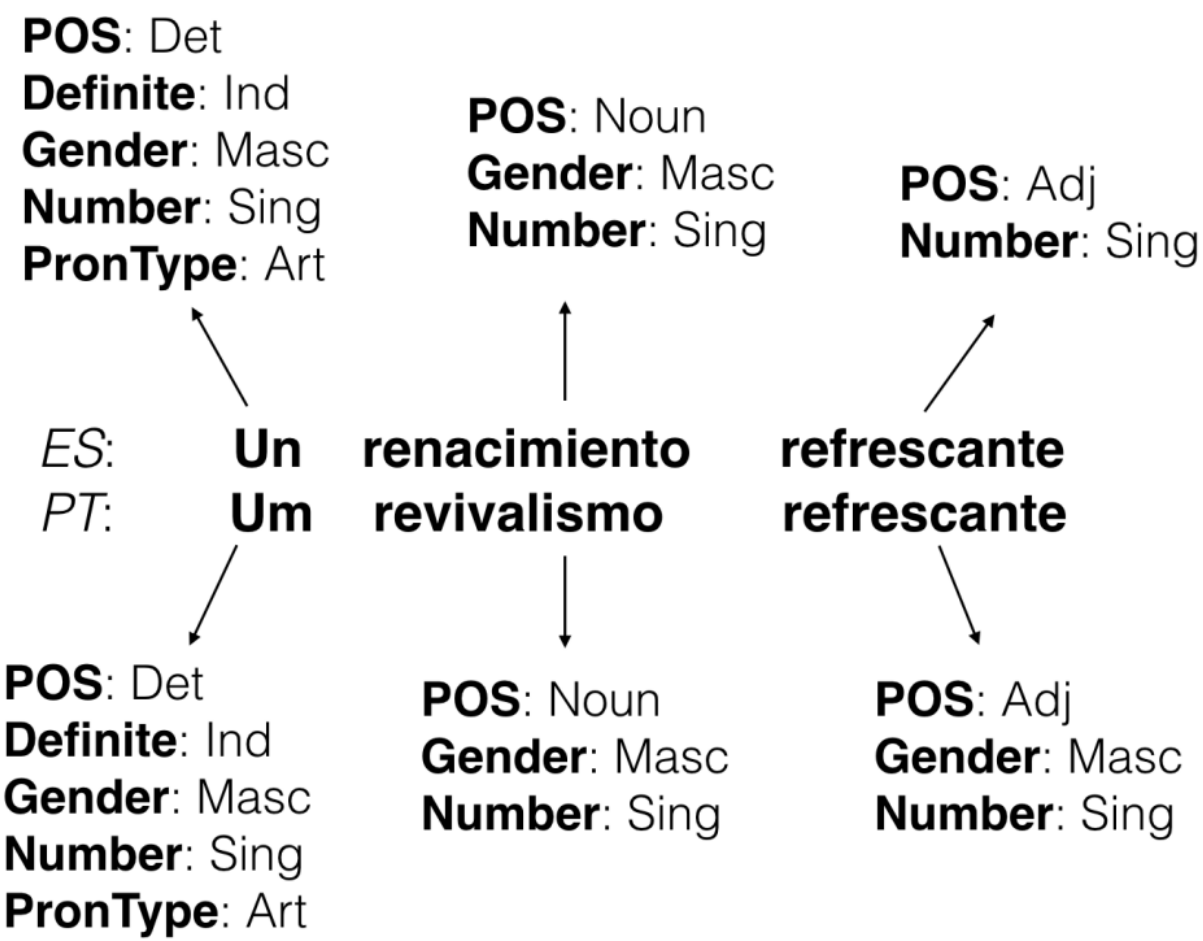

Figure: Tag morphologique d'une phrase portugaise et sa traduction en espagnol.

Le dataset

Nous avons utilisé le dataset Universal Dependencies 2.13.
Le dataset en français contient :

- 47498 phrases
- 849476 mots
- 76048 mots uniques

On a recensé :

- 19 classes *pos*.
- 28 classes *morphy*, avec un nombre de possibilités entre 2 et 13

Les *morphy*


```
"_": ["Not", "Yes"],
"<PAD>": ["Not", "Yes"],
"Emph": ["Not", "No", "Yes"],
"Gender": ["Not", "Fem", "Masc", "Neut", "Fem,Masc"],
"Number": ["Not", "Sing", "Plur"],
"Person": ["Not", "3", "1", "2"],
"PronType": ["Not", "Prs", "Dem", "Ind", "Int", "Rel", "Art", "Emp", "Tot", "Neg
"Typo": ["Not", "Yes"],
"ExtPos": ["Not", "VERB", "PROPN", "NOUN", "ADV", "ADP", "ADJ", "SCONJ", "DET",
"Tense": ["Not", "Pres", "Imp", "Past", "Fut"],
"VerbForm": ["Not", "Ger", "Fin", "Part", "Inf"],
"NumType": ["Not", "Ord", "Card", "Mult", "Frac"],
"Mood": ["Not", "Ind", "Imp", "Sub", "Cnd"],
"Poss": ["Not", "Yes"],
"Case": ["Not", "Acc", "Nom", "Gen"],
"Reflex": ["Not", "Yes"],
"Polarity": ["Not", "Neg", "Pos"],
"Degree": ["Not", "Cmp", "Pos", "Sup"],
"Style": ["Not", "Vrnc", "Expr", "Arch", "Slng", "Coll"],
"Number[psor]": ["Not", "Plur", "Sing"],
"Person[psor]": ["Not", "1", "3", "2"],
"NumForm": ["Not", "Roman", "Word", "Combi", "Digit"],
"Abbr": ["Not", "Yes"],
"Voice": ["Not", "Pass", "Act"],
"AdpType": ["Not", "Prep"],
"Foreign": ["Not", "Yes"],
"Definite": ["Not", "Ind", "Def"],
"Morph": ["Not", "VFin", "VInf", "VPar"]
```

création de séquences

Idée

On rajoute des caractères ⟨PAD⟩ pour compléter les phrases, on rajoute aussi un label qui correspond au pad.

Si on fixe une longueur de séquence à 5, on va transformer la phrase :

type	**exemple**	**longueur**
phrase	Les poissons sont des animaux vertébrés .	7
Séquence	Les poissons sont des animaux vertébrés . ⟨PAD⟩ ⟨PAD⟩ ⟨PAD⟩	5 5

Table: Exemple de la création du séquences

La gestion des mots inconnus

Idée

Si le modèle rencontre un mot inconnus, il va le remplacer par le mot $\langle \text{UNK} \rangle$. Le modèle doit alors aussi apprendre ce mot lors de l'entraînement.

- Avant l'entraînement, on fait un dropout des mots avec un taux d'oublie de 1%.
- On obtient un nombre de mots dans le vocabulaire de taille 67814 (contre 76048 dans le dataset).
- On a fixé la seed avant

Encodage des labels

Pour les *pos* :

- On remplace le label par son indice dans la liste des labels.

Pour les *morphy* :

- On remplace le label par une liste d'indice, où l'élément i de la liste correspond à l'indice du i-ème type de morphy.

Exemple : *Emph=No|Number=Sing|Person=1|PronType=Prs*
[0, 0, 1, 0, 1, 2, 9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Shape

- Entrée : x de taille $B \times K$ contenant les indices des mots
- Sortie de *pos*: y_{pos} de taille $B \times K \times 19$
- Sortie de *morphy*: y_{morphy} de taille $B \times K \times 28 \times 13$

Modèle *GET_POS*

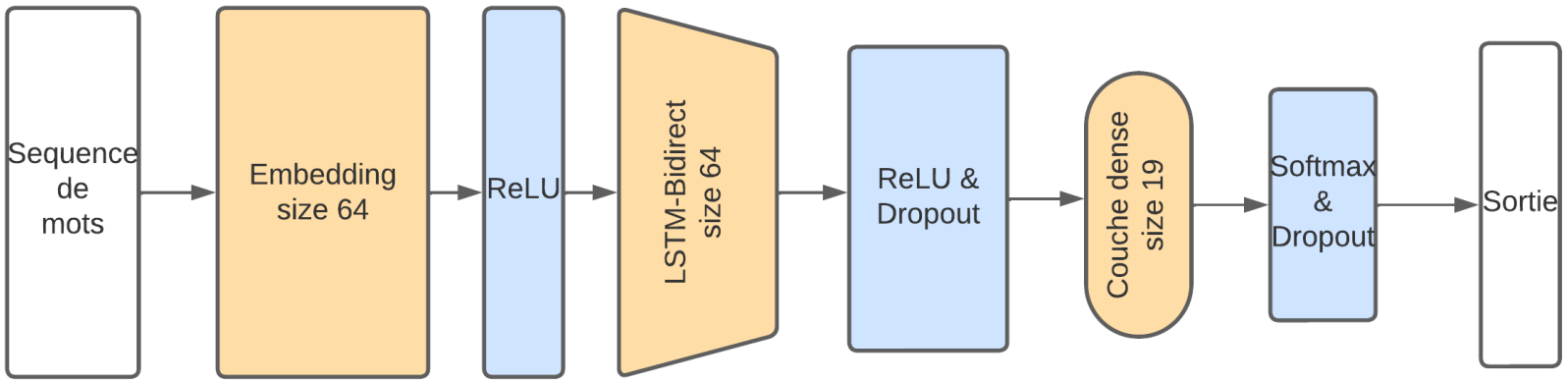

Figure: Modèle *GET_POS*

Modèle *SUPERTAG*

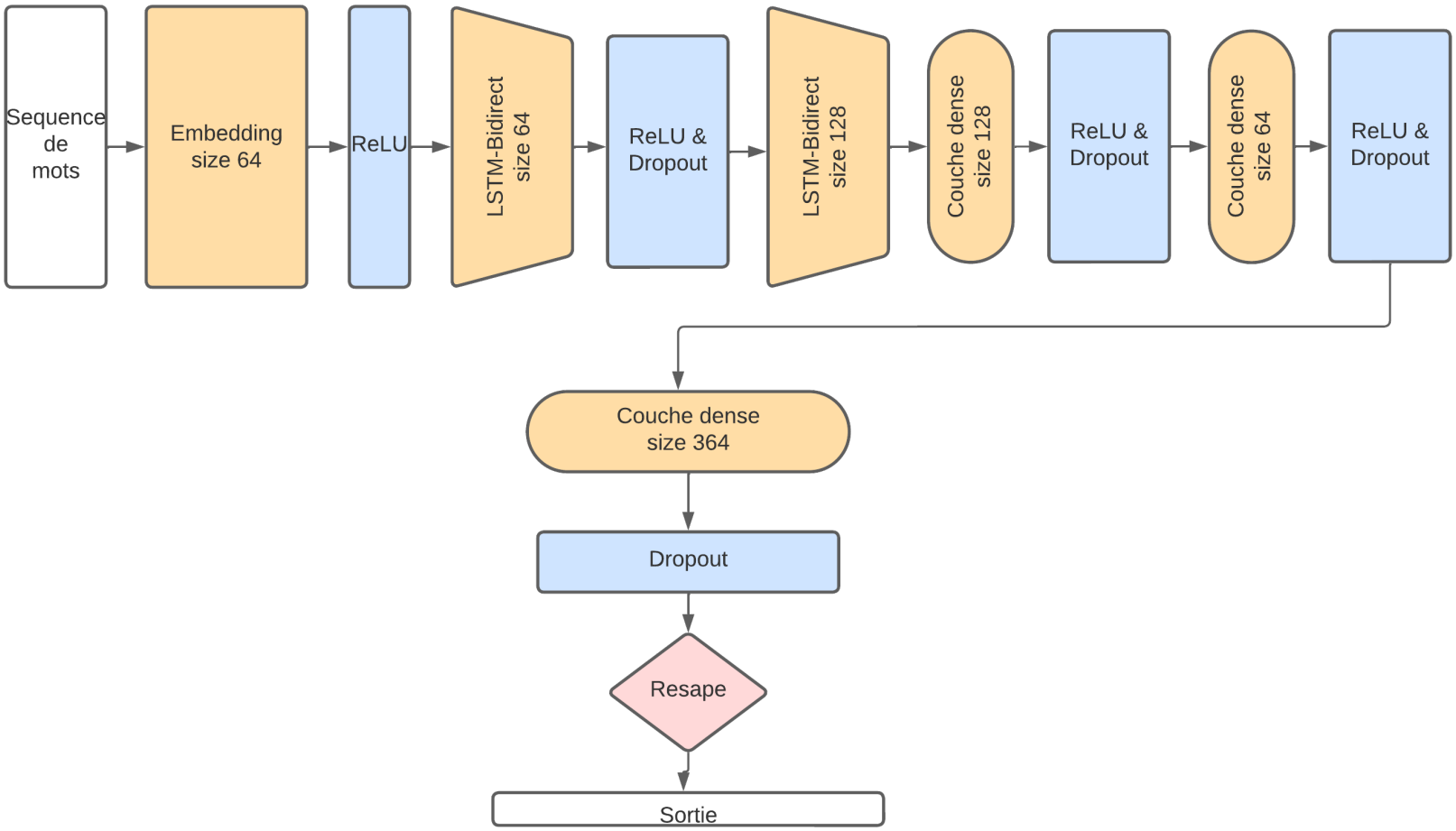

Figure: Modèle *SUPERTAG*

Modèle *SEPARATE*

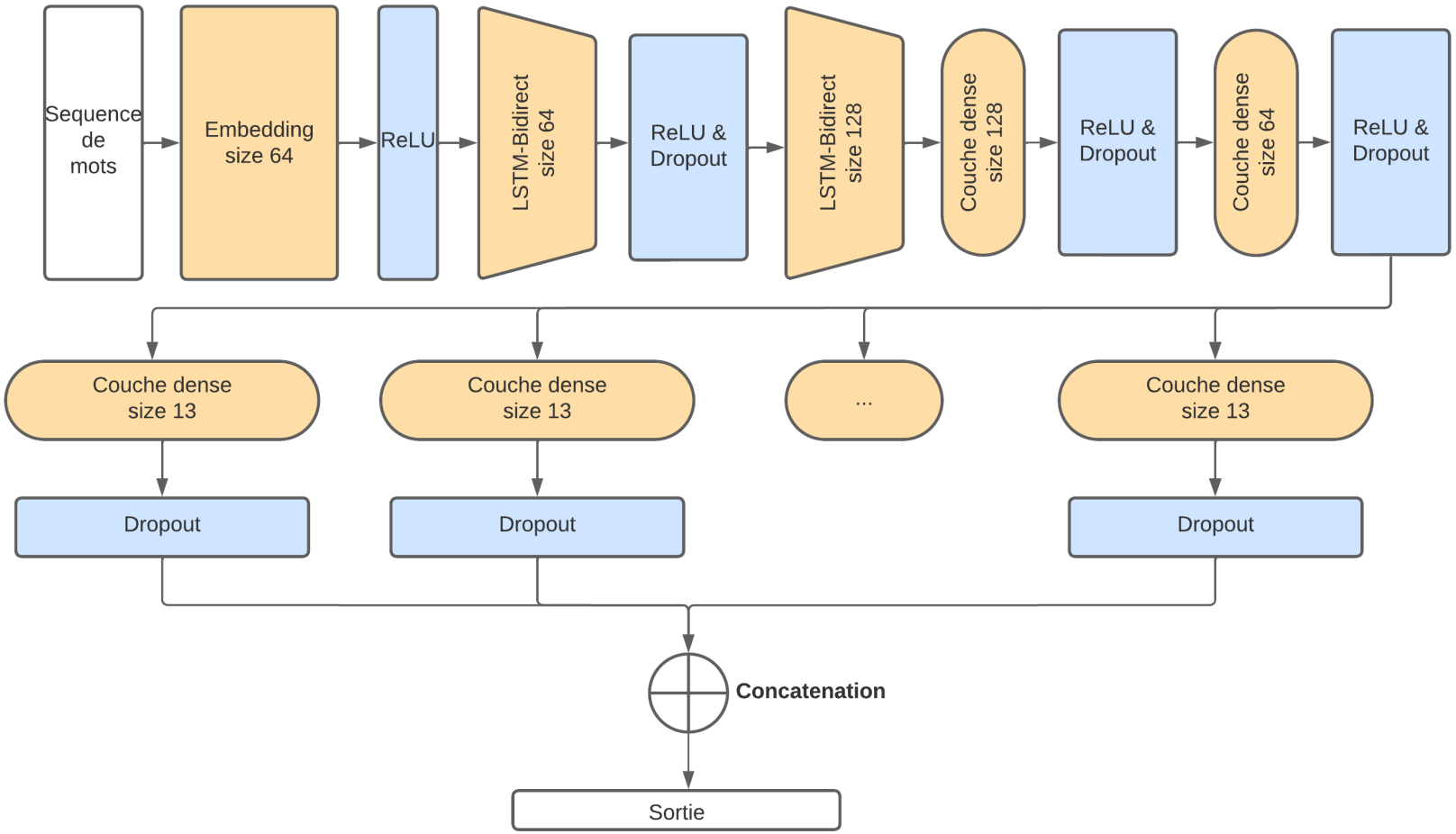

Figure: Modèle *SEPARATE*.

Modèle *FUSION*

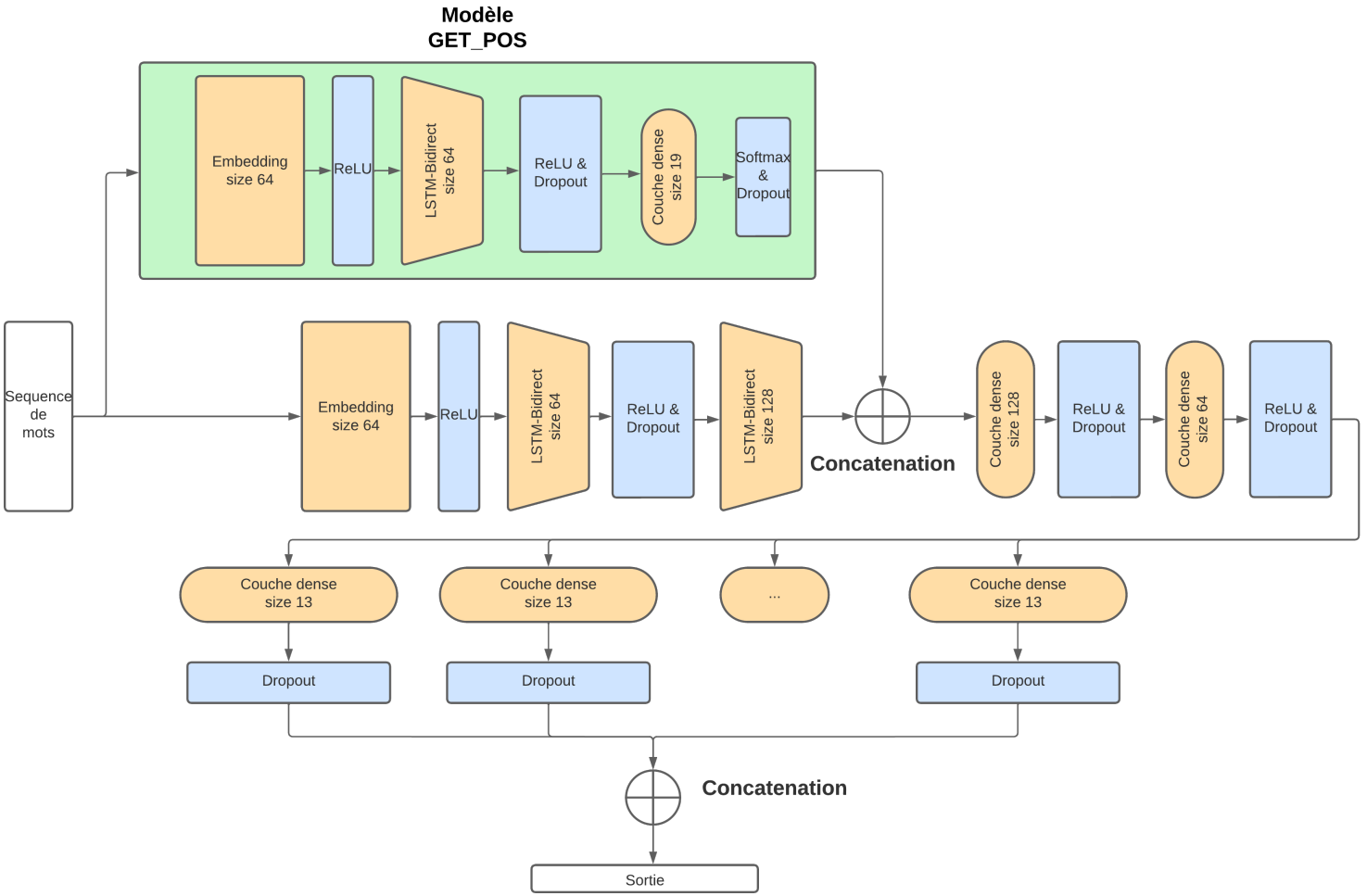

Figure: Modèle *FUSION*

Loss et Métriques

Les loss :

- crossentropy pour le *pos*
- moyenne de la crossentropy sur les 28 classes pour le *morphy*

Les métriques :

- accuracy micro
- accuracy macro (pour le *pos*)
- allgood (pour le *morphy*)

Résultats : *GET_POS*

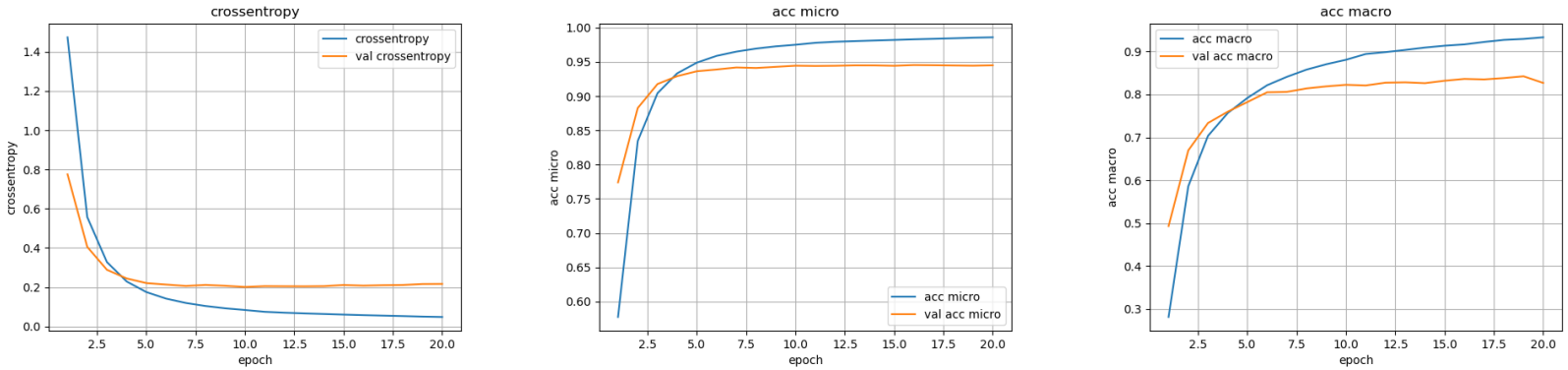

Figure: Entraînement du modèle *GET_POS*.

Nom du modèle	crossentropy	accuracy micro	accuracy macro
GET_POS	0.204	0.944	0.816

Table: Résultats du modèle *GET_POS* sur la base de données de teste.

Résultats des modèles *SUPERTAG* et *SEPARATE*

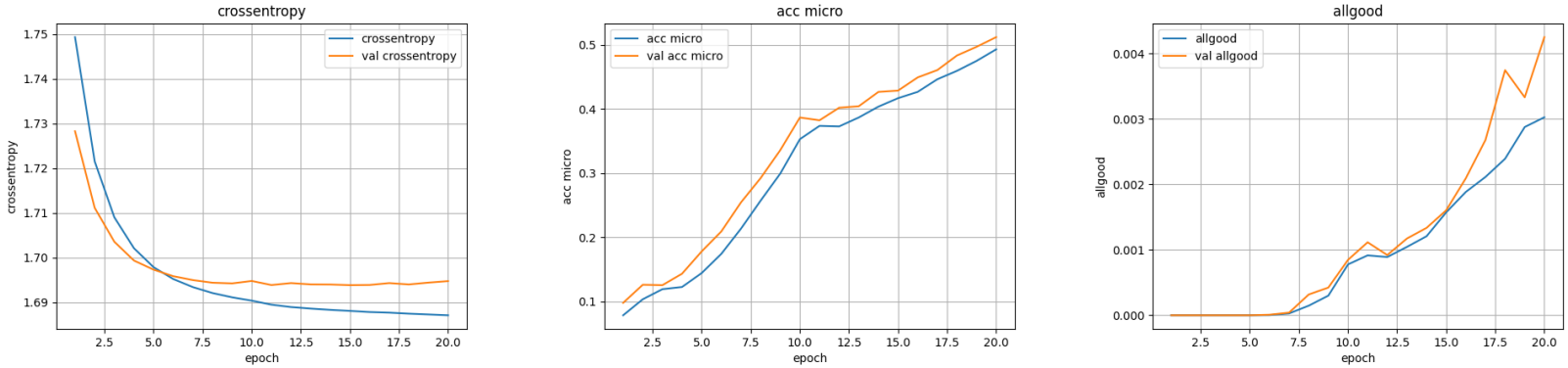

Figure: Entraînement du modèle *SUPERTAG*.

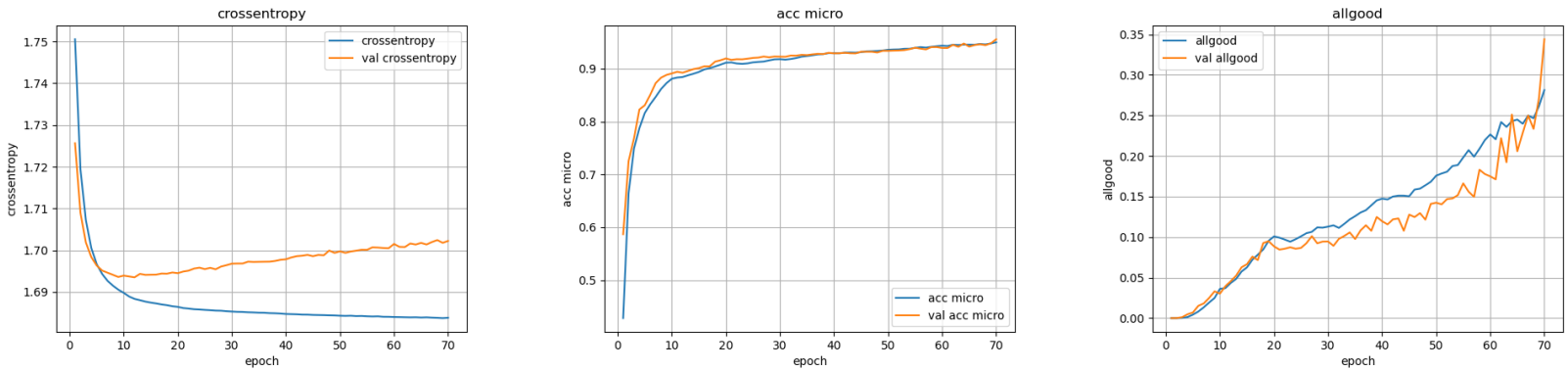

Figure: Entraînement du modèle *SEPARATE*.

Résultats du modèle *FUSION* et résultats de teste

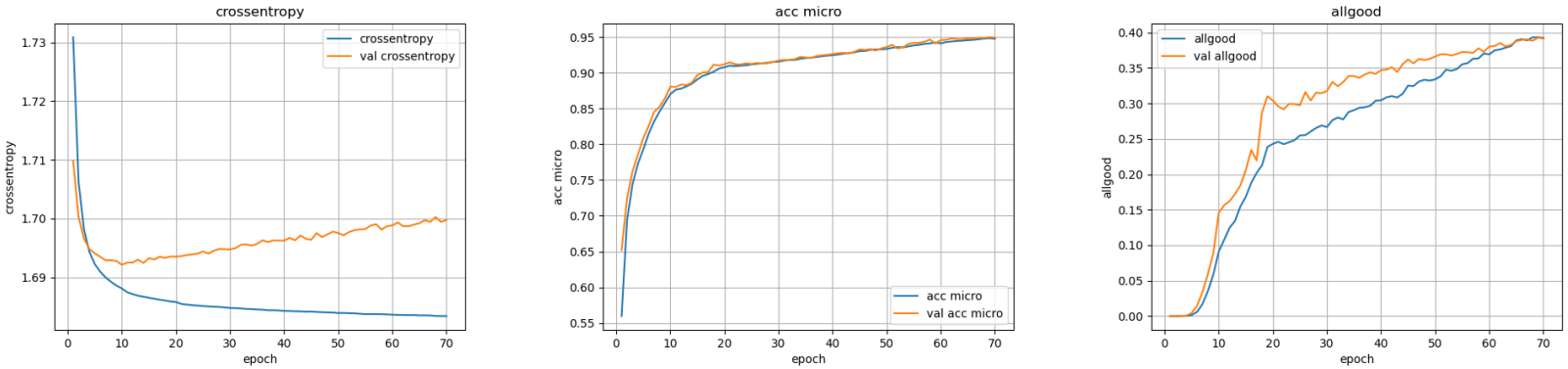

Figure: Entraînement du modèle *FUSION*.

Nom du modèle	crossentropy	accuracy micro	all good
BASELINE	-	0.980	0.791
SUPERTAG	1.700	0.436	0.002
SEPARATE	1.70	0.893	0.046
FUSION	1.698	0.884	0.154

Table: Résultats de test sur la prédiction des *morphy*

Inférences

Inférences sur la phrase : *Les bananes sont jaunes et mûrs.*
un mot inconnu : *mûrs*

modèle	**inférences**
BASELINE	Gender=Fem|Number=Plur
SUPERTAG	NumForm=Roman|Abbr=Yes|Morph=VInf
SEPARATE	Gender=Neut|Number=Plur|VerbForm=Ger Degree=Pos|Abbr=Yes
FUSION	Gender=Fem,Masc|Number=Plur|PronType=Tot NumType=Mult|Case=Nom|NumForm=Combi

Table: Inférences du mot *bananes*

Inférences

modèle	**inférences**
BASELINE	⟨PAD⟩=Yes
SUPERTAG	PronType=Emp|Style=Coll
SEPARATE	⟨PAD⟩=Yes|PronType=Int,Rel|Case=Nom Degree=Sup|NumForm=Combi
FUSION	⟨PAD⟩=Yes|Gender=Neut|PronType=Int,Rel NumType=Mult|Degree=Sup|Style=SIng NumForm=Combi

Table: Inférences du dernier⟨PAD⟩.

Nos difficultés rencontrées

- Trouver la bonne loss pour les *morphy*. Car avant, on avait seulement la crossentropy et 0.02% d'accuracy.
- Dans le modèle *SEPARATE*, on a l'ensemble des dernières couches Dense dans une liste. Et donc le *model.parameters()* ne donnent pas les paramètres de ces couches.
- Avant, on n'avait pas ajouté les $-\infty$ dans le modèle *SEPARATE*, donc on avait de moins bonnes performances

Les pistes d'améliorations

- Changer la façon dont on crée les séquences.
- Ne pas prendre en comte les $\langle \text{PAD} \rangle$ dans la loss et les métriques.
- Changer la façon de mettre des mots inconnus. Car là, les mots inconnus ne changent pas sur les epochs.
- comparer les résultats en fonction de la taille des séquences
- Mettre une couche Dense après les couches separate.
- Remplacer les modules LSTM par des Transformers.