Séminaire LIS

Titre : End-to-end model for named entity recognition from speech without paired training data

Salima Mdhaffar

Post-doc au Laboratoire d'Informatique d'Avignon (LIA)

LABORATOIRE INFORMATIQUE D'AVIGNON

AVIGNON UNIVERSITÉ

Introduction

- **Avancements significatives en traitement automatique de la parole**
  - Architecture neuronale de bout en bout
  - Apprentissage auto-supervisée
  - Existence d’une large quantité de données

Introduction

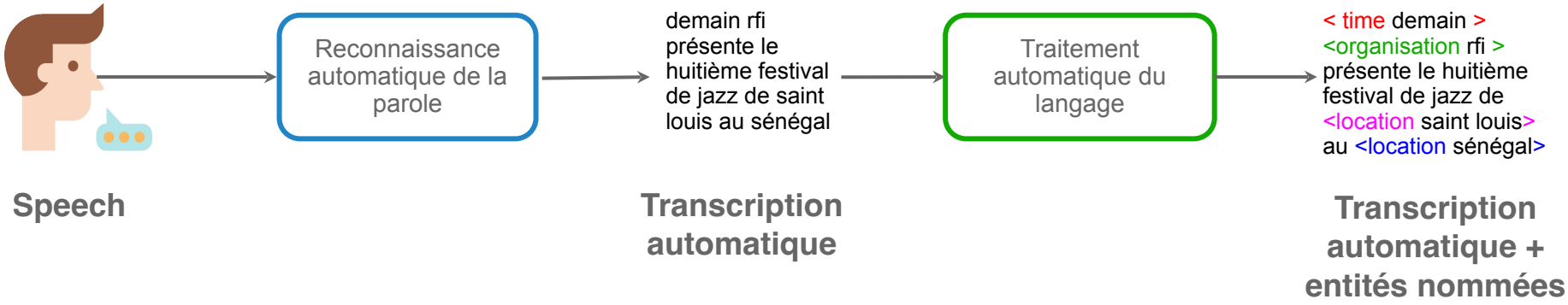
- **Avancements significatives en traitement automatique de la parole**
  - Architecture neuronale de bout en bout
  - Apprentissage auto-supervisée
  - Existence d'une large quantité de données

- **Reconnaissance d'entités nommées**
  - Tâche de la compréhension de la parole
  - Détecter les entités nommées dans la parole
  - **Entité nommée** : brique élémentaire de l'information contenue dans les documents

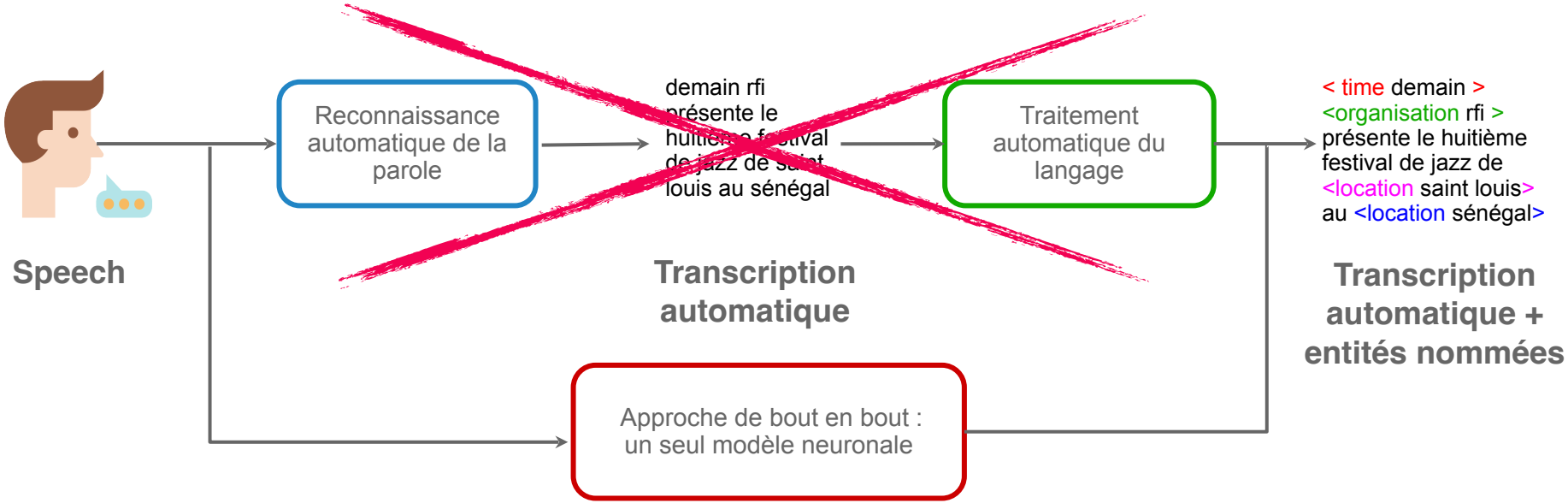
<time demain > <organisation rfi > présente le huitième festival de jazz de <location saint louis > au <location sénégal >

Introduction

Approche cascade



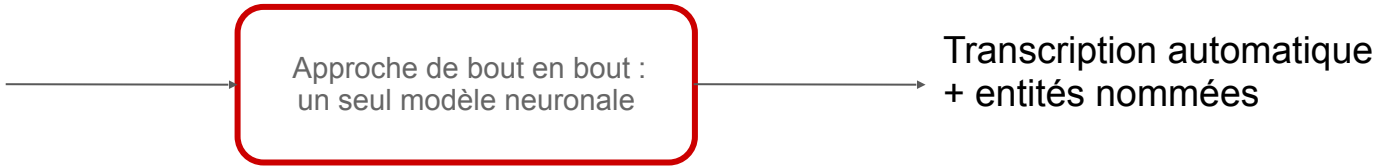
Approche bout en bout



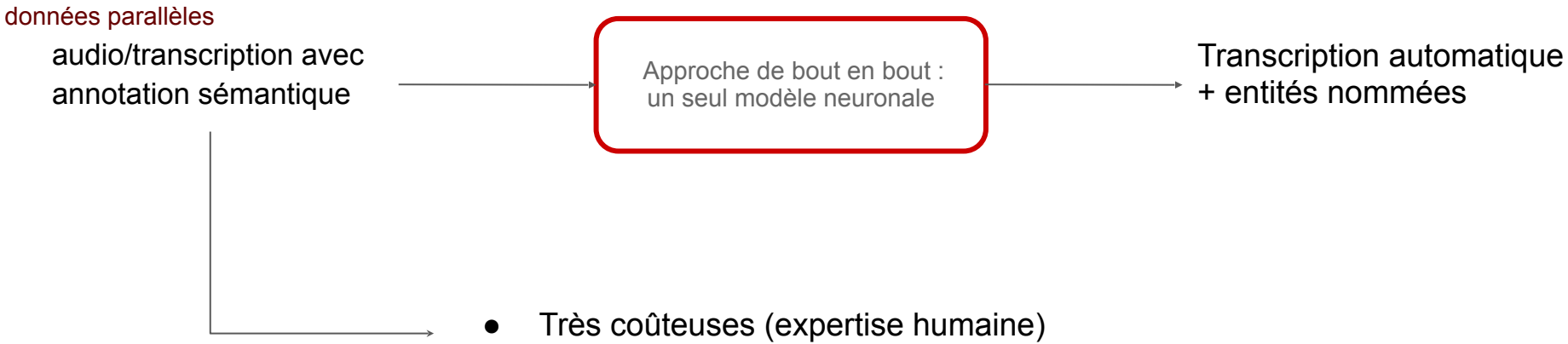
Approche bout en bout

données parallèles

audio/transcription avec annotation sémantique



Approche bout en bout



- Très coûteuses (expertise humaine)
- Indisponible ou limité pour les langues peu dotées

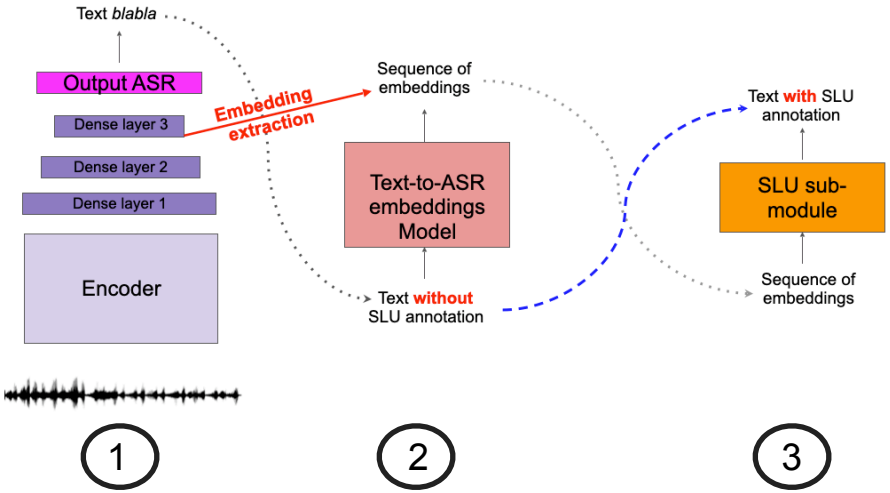
Introduction

Proposer une approche neuronale de bout en bout pour construire un modèle permettant l'extraction sémantique pour un scénario où zéro données parallèles (audio/transcription avec annotation sémantique) est disponible.

Approche proposée

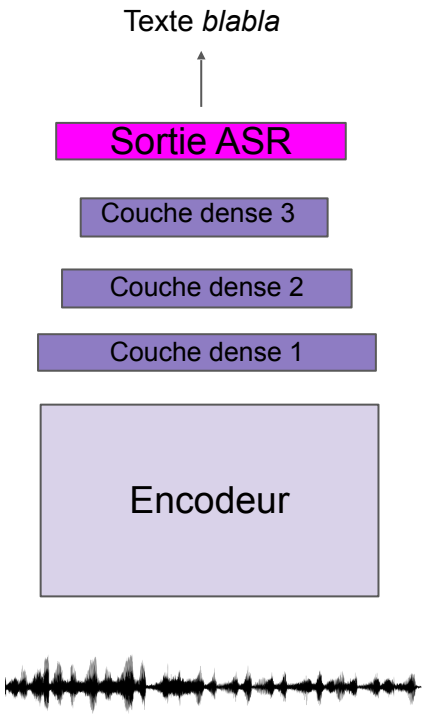
L'approche proposée est composé de trois modèles :

1. Modèle Text-to-ASR Embeddings
2. Modèle SLU
3. Modèle final



Approche proposée

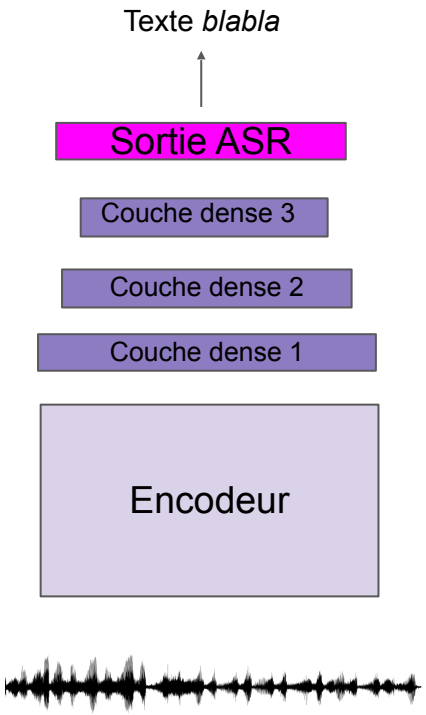
1. text-to-ASR embeddings | 2. Modèle SLU | 3. Modèle finale



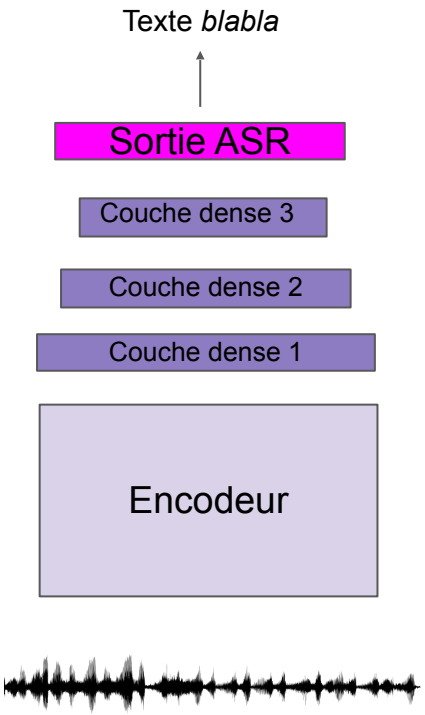
Données textuelles avec annotation en entités nommées

Approche proposée

1. text-to-ASR embeddings 2. Modèle SLU 3. Modèle finale



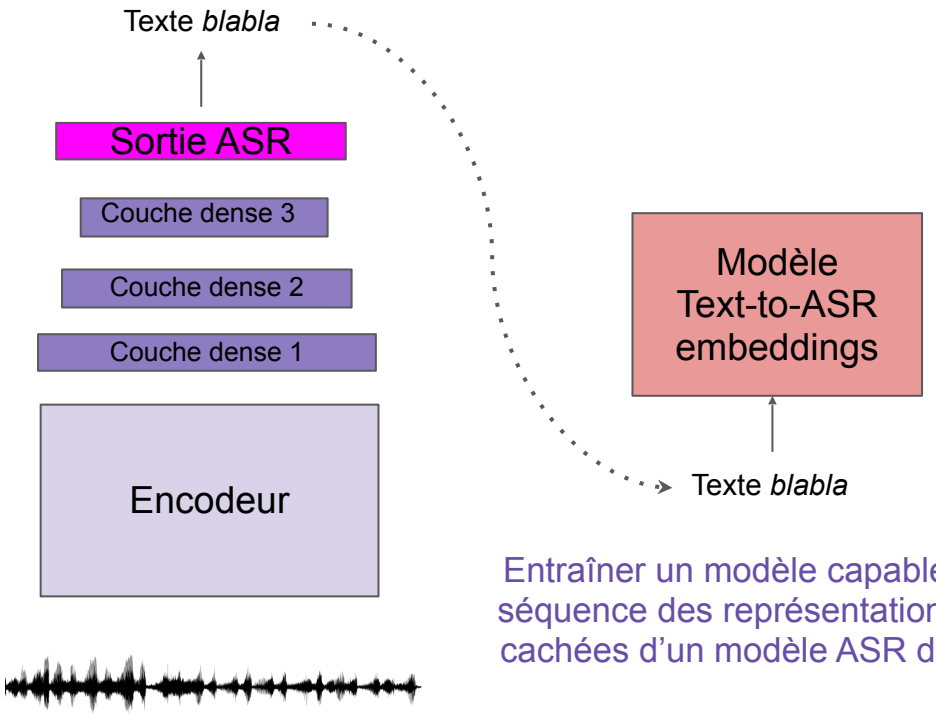
Approche proposée



Modèle Text-to-ASR embeddings

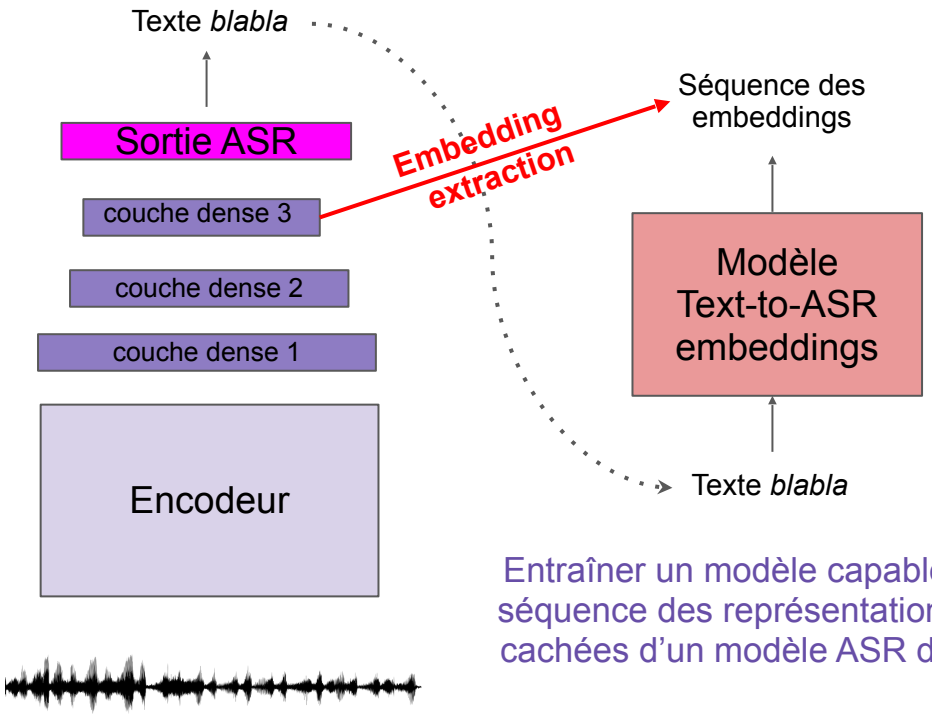
Entraîner un modèle capable de simuler la séquence des représentations/embeddings cachées d'un modèle ASR de bout en bout

Approche proposée



Entraîner un modèle capable de simuler la séquence des représentations/embeddings cachées d'un modèle ASR de bout en bout

Approche proposée



Entraîner un modèle capable de simuler la séquence des représentations/embeddings cachées d'un modèle ASR de bout en bout

Approche proposée

Données textuelles avec annotation en entités nommées

Approche proposée

Données textuelles avec annotation en entités nommées

Texte **sans** annotation SLU

Modèle Text-to-ASR embeddings

Extraire les embeddings pour les données textuelles annotés en entités nommées

Approche proposée

Données textuelles avec annotation en entités nommées

Texte **sans** annotation SLU

Modèle Text-to-ASR embeddings

Séquence des embeddings

Extraire les embeddings pour les données textuelles annotés en entités nommées

Approche proposée

Données textuelles avec annotation en entités nommées

Texte **sans** annotation SLU

Modèle Text-to-ASR embeddings

Séquence des embeddings

Extraire les embeddings pour les données textuelles annotés en entités nommées

Module NLU

Entraîner un modèle SLU à partir de la séquence des embeddings générée automatiquement

Approche proposée

Données textuelles avec annotation en entités nommées

Texte **sans** annotation SLU

Modèle Text-to-ASR embeddings

Séquence des embeddings

Séquence des embeddings

Module NLU

Extraire les embeddings pour les données textuelles annotés en entités nommées

Entraîner un modèle SLU à partir de la séquence des embeddings générée automatiquement

Approche proposée

Données textuelles avec annotation en entités nommées

Texte **sans** annotation SLU

Modèle Text-to-ASR embeddings

Séquence des embeddings

Extraire les embeddings pour les données textuelles annotés en entités nommées

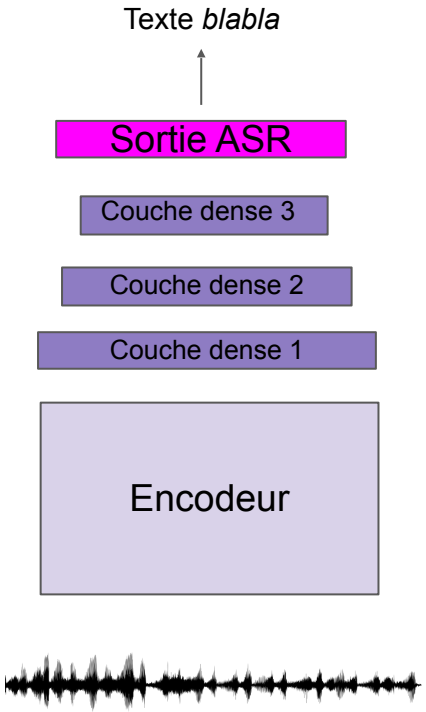
Texte **avec** annotation SLU

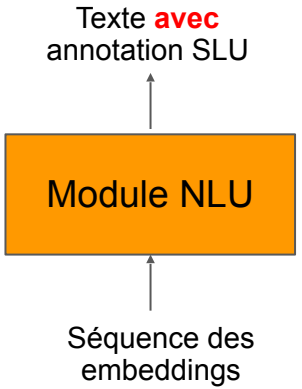
Module NLU

Séquence des embeddings

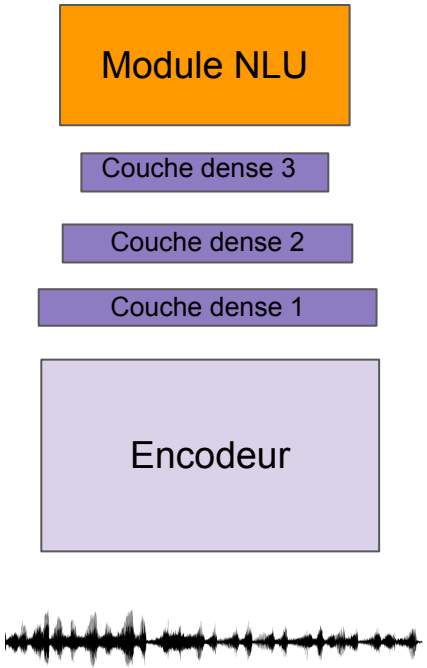
Entraîner un modèle SLU à partir de la séquence des embeddings générée automatiquement

Approche proposée

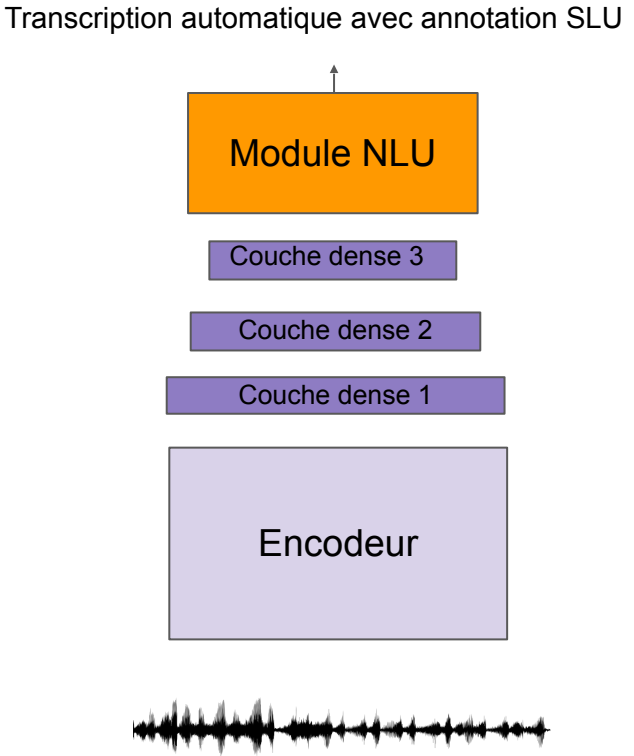




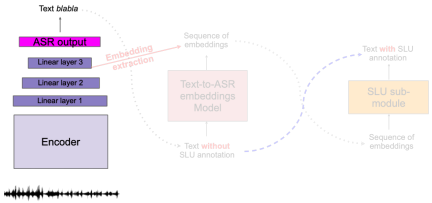
Approche proposée



Approche proposée



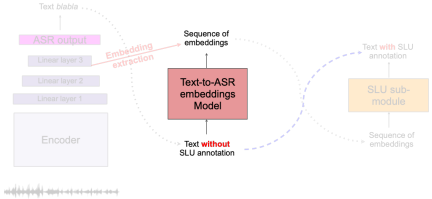
Cadre expérimental : données



- Données d'actualités collectés à partir du radio et de la télévision (Français):
  - REPERE [Giraudel et al, 2012] : 35,5 heures
  - ESTER2 [Galliano et al, 2009] : 91 heures
  - EPAC [Grouin et al, 2011] : 70,5 heures

=> total : 197 heures

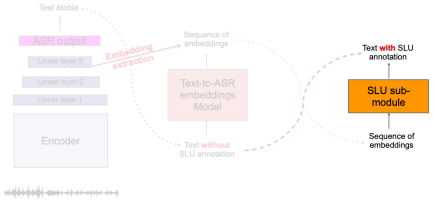
Cadre expérimental : données



Entrée	Labels
Texte *blabla* →	0.6 −0.2 0.8 0.9 −0.1 −0.9 −0.7
Texte *blabla* →	0.7 0.3 0.9 −0.7 0.1 −0.5 −0.4
.	.
.	.
Texte *blabla* →	0.5 −0.4 0.7 0.8 0.9 −0.7 −0.6

- Entrée : transcription automatique pour 197 heures utilisées dans l'apprentissage de l'ASR
  - WER pour la transcription automatique : 7.41%
- Sortie : embeddings pour 197 heures utilisées dans l'apprentissage de l'ASR

Cadre expérimental : données



- Train : transcription manuelle de QUAERO [Grouin et al, 2011]
  - 7 types d'entités nommés : personne, localisation, organization, produit, temps, quantité et fonction.
  - Exemple :

(a) demain rfi présente le huitième festival de jazz de saint louis au sénégal

(b) <time demain > <organisation rfi > présente le huitième festival de jazz de <location saint louis > au <location sénégal >

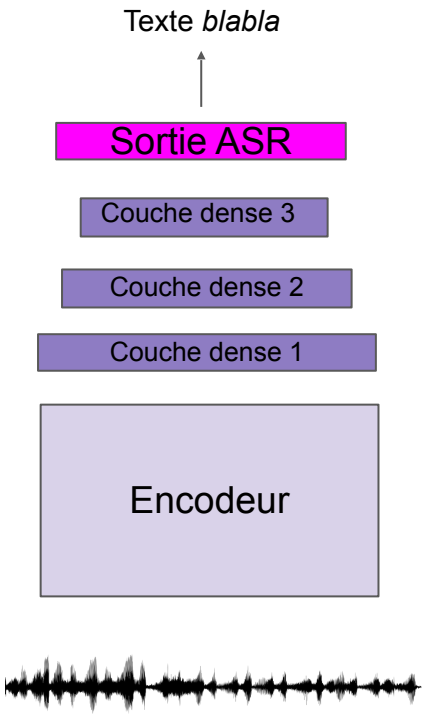
Cadre expérimental : données

- Dev & test : audio / transcription avec annotation sémantique

Data	# mots	# vocab	# Concepts	# Utterance	# Heures
Train	993.7k	33.5k	89.6k	90.3k	81h
Dev	81.5k	8.7k	5.5k	10.9k	6.5h
Test	119.3k	13.4k	10.8k	4.0k	10h

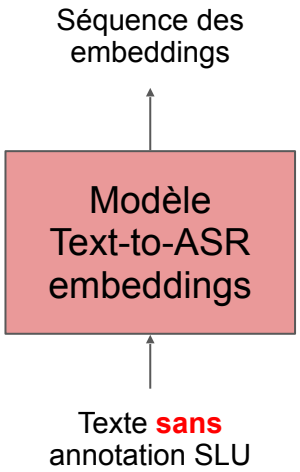
Statistiques des données QUAERO

Cadre expérimental : architecture



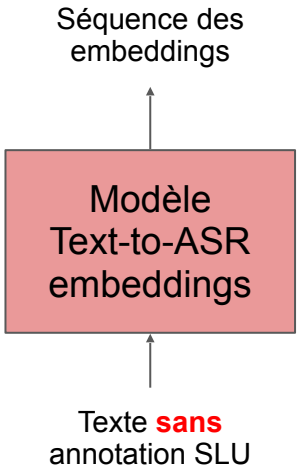
- Encodeur : wav2vec 2.0 LeBenchmark [Evain et al, 2021]
- Couches dense:
  - Couche dense 1 : 1024
  - Couche dense 2 : 512
  - Couche dense 3 : 80
- Sortie ASR : taille 47 / fonction du loss : ctc

Cadre expérimental : architecture



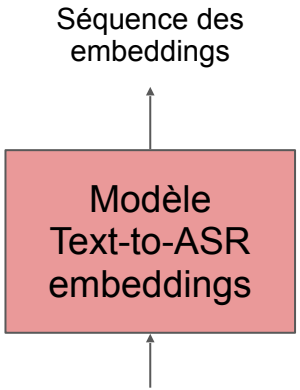
- basé sur l'architecture Tacotron 2 [Shen et al, 2018]
  - module qui transforme les embeddings de caractères en des spectrogrammes
  - vocodeur qui transforme les spectrogrammes en fichier audio

Cadre expérimental : architecture

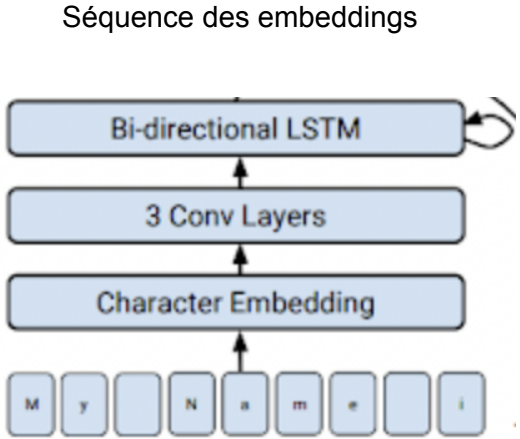


- basé sur l'architecture Tacotron 2 [Shen et al, 2018]
  - **module qui transforme les embeddings de caractères en des spectrogrammes**
  - vocodeur qui transforme les spectrogrammes en fichier audio

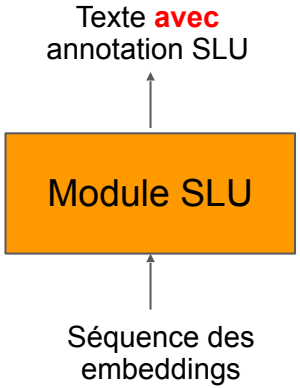
Cadre expérimental : architecture



- basé sur l'architecture Tacotron 2 [Shen et al, 2018]
  - **module qui transforme les embeddings de caractères en des spectrogrammes**
  - vocodeur qui transforme les spectrogrammes en fichier audio

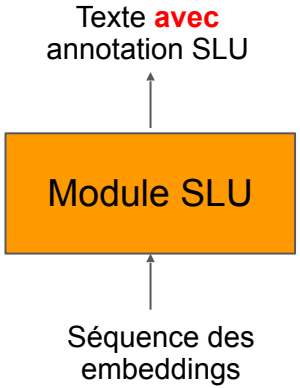


Cadre expérimental : architecture



- BiLSTM architecture:
  - 5 stack of BiLSTM couches
  - Dimension : 512
- Sortie : taille de softmax = 54 (47 ASR + 7 entities)

Cadre expérimental : architecture



- BiLSTM architecture:
  - 5 stack of BiLSTM couches
  - Dimension : 512
- Sortie : taille de softmax = 54 (47 ASR + 7 entities)

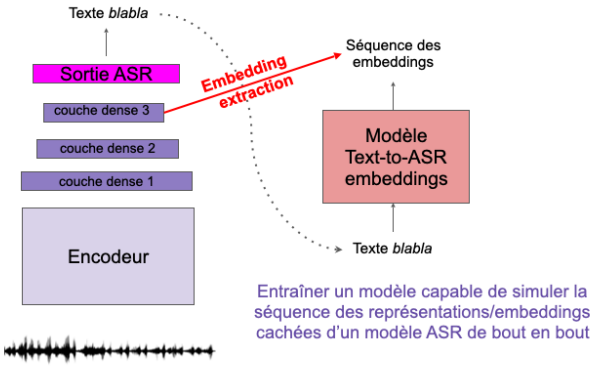
A
B
.
.
.
Z

personne
organisation
fonction
localisation
produit
temps
quantité

Results

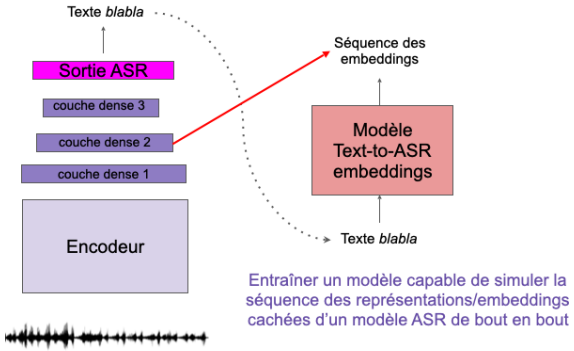
- Évaluation de la qualité reconnaissance de la parole en utilisant les embeddings générés
  - (1) Apprendre des modèles Text-to-Embeddings à partir de différentes couches

Couche dense 3



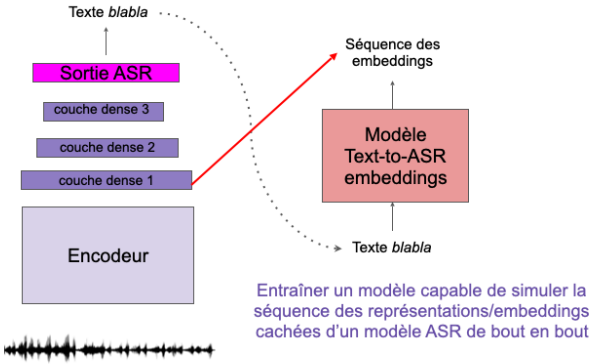
Entraîner un modèle capable de simuler la séquence des représentations/embeddings cachées d'un modèle ASR de bout en bout

Couche dense 2



Entraîner un modèle capable de simuler la séquence des représentations/embeddings cachées d'un modèle ASR de bout en bout

Couche dense 1

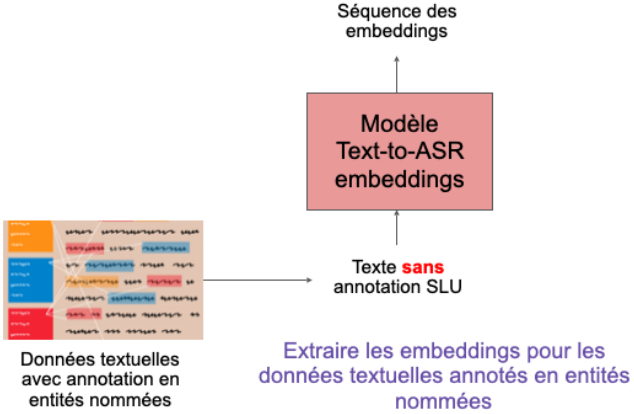


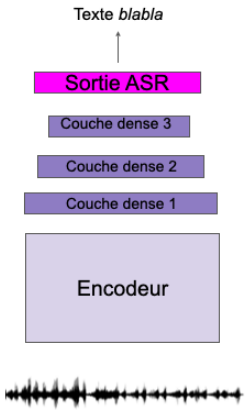
Entraîner un modèle capable de simuler la séquence des représentations/embeddings cachées d'un modèle ASR de bout en bout

Results

- Évaluation de la qualité reconnaissance de la parole en utilisant les embeddings générés
  - (1) Apprendre des modèles Text-to-Embeddings à partir de différentes couches
  - (2) Pour chaque modèle, générer des embeddings avec les modèles appris et remplacer la couche ciblée avec embeddings générés (en inférence)

QUAERO Train textual dataset

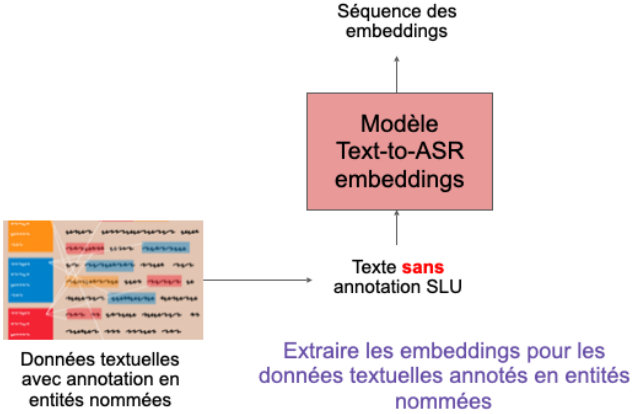


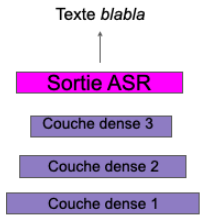


Results

- Évaluation de la qualité reconnaissance de la parole en utilisant les embeddings générés
  - (1) Apprendre des modèles Text-to-Embeddings à partir de différentes couches
  - (2) Pour chaque modèle, générer des embeddings avec les modèles appris et remplacer la couche ciblée avec embeddings générés (en inférence)

QUAERO Train textual dataset

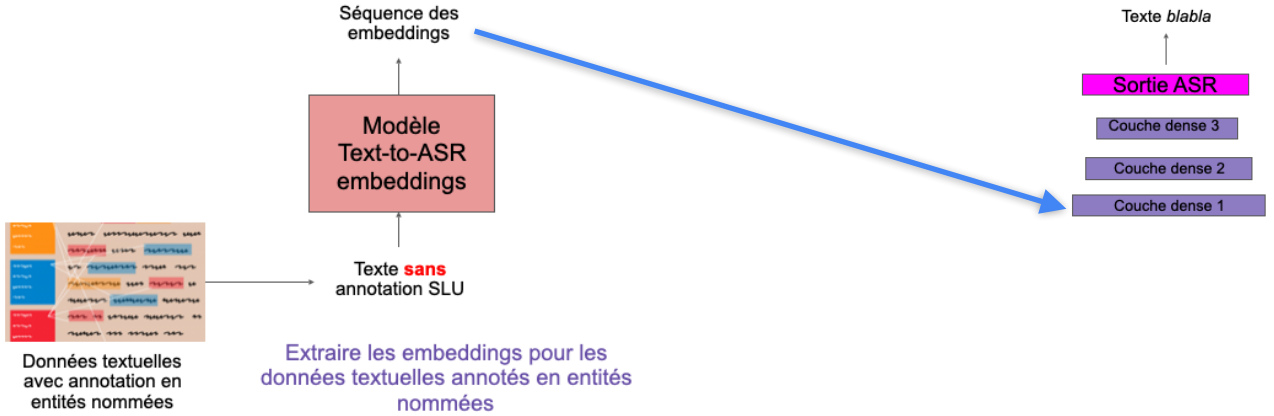




Results

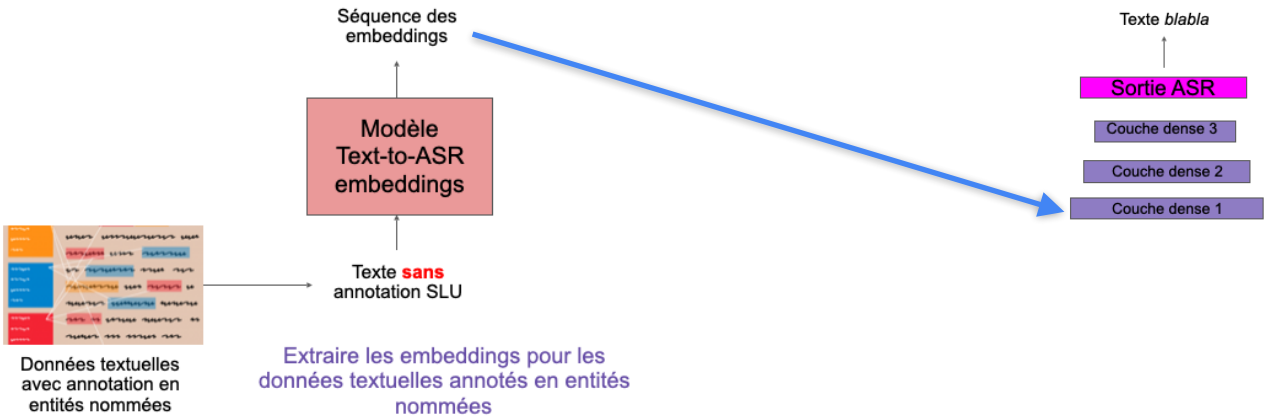
- Évaluation de la qualité reconnaissance de la parole en utilisant les embeddings générés
  - (1) Apprendre des modèles Text-to-Embeddings à partir de différentes couches
  - (2) Pour chaque modèle, générer des embeddings avec les modèles appris et remplacer la couche ciblée avec embeddings générés (en inférence)

QUAERO Train textual dataset



Results

- Évaluation de la qualité reconnaissance de la parole en utilisant les embeddings générés
  - (1) Apprendre des modèles Text-to-Embeddings à partir de différentes couches
  - (2) Pour chaque modèle, générer des embeddings avec les modèles appris et remplacer la couche ciblée avec embeddings générés (en inférence)
  - (3) Calculer un WER



Results

- Évaluation de la qualité reconnaissance de la parole en utilisant les embeddings générés
  - (1) Apprendre des modèles Text-to-Embeddings à partir de différentes couches
  - (2) Pour chaque modèle, générer des embeddings avec les modèles appris et remplacer la couche ciblée avec embeddings générés (en inférence)
  - (3) Calculer un WER

	Layer 1	Layer 2	Layer 3	Wav QUAERO
WER	80,29 %	56,68 %	15,56 %	9,08 %

WER pour les embeddings générés sur les données de train QUAERO

Results

- **Métrique d'évaluation:**
  - **NEER: N**amed **E**ntity **E**rror **R**ate

$$\text{NEER} = \frac{S + D + I}{n}$$

S: Nombre d'erreurs de substitution
D: Nombre d'erreurs de suppression
I: Nombre d'erreurs d'insertion
n: Nombre de références

Results

- **Métrique d'évaluation:**
  - **NEER: N**amed **E**ntity **E**rror **R**ate

$$\text{NEER} = \frac{S + D + I}{n}$$

S: Nombre d'erreurs de substitution
D: Nombre d'erreurs de suppression
I: Nombre d'erreurs d'insertion
n: Nombre de références

Référence: < time demain > <organisation rfi > présente le huitième festival de jazz de <location saint louis> au <location sénégal>
Hypothèse: < time demain > <organisation rfi > présente le huitième festival de jazz de <person saint louis> au sénégal

NEER evaluation:
Référence: <time> <organisation> <location> <location>
Hypothèse: <time> <organisation> <person>
= = S D

NEER = 2/4 = 0.5

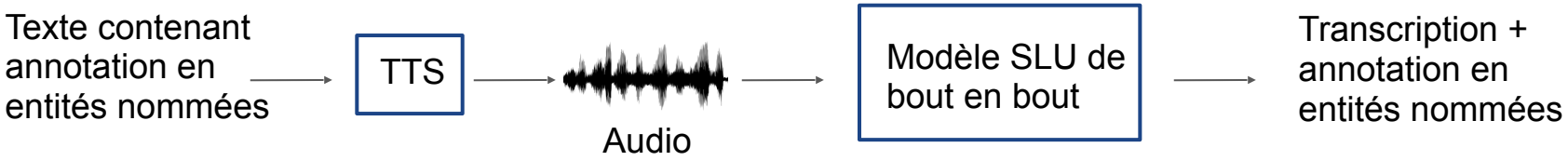
Results

Données d'apprentissage	Dev NEER	Test NEER
ASR embeddings simulation (ours)	**47,6 %**	**39,1 %**
Oracle (real data)	45,9 %	34,1 %

Évaluation du modèle end-to-end NER sans des données d'apprentissage parallèles en comparaison avec une approche miracle où des données parallèles sont disponibles

Results

Comparer l'approche proposée avec un système appris la voix synthétique :



	Wav TTS	Wav QUAERO
WER	21,60 %	9,08 %

Evaluation des données train QUAERO générés avec le TTS

Results

Données d'apprentissage	Dev NEER	Test NEER
ASR embeddings simulation (ours)	**47,6 %**	**39,1 %**
Synthetic speech (all weights are updated)	65,2 %	62,7 %
Synthetic speech (frozen speech encoder)	86,4 %	92,5 %
Oracle (real data)	45,9 %	34,1 %

Évaluation de l'approche proposée pour entraîner un modèle end-to-end sans des données parallèles en comparaison avec d'autres approches en utilisant de la voix synthétique

Results

Comparer l'approche proposée avec un système en cascade :

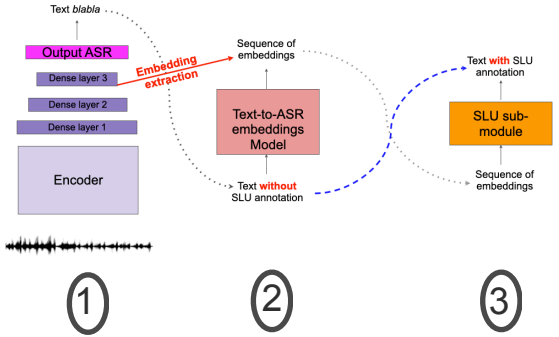
Données d'apprentissage	Dev NEER	Test NEER
ASR embeddings simulation (ours)	**47,6 %**	**39,1 %**
Text-to-text NER on manual transcripts	39,6 %	30,0 %
Text-to-text NER on automatic transcripts	48,0 %	40,2 %

Résultats de notre approche en comparaison de en comparaison d'une simple approche en cascade

Conclusion

- Une approche neuronale de bout en bout pour construire un modèle permettant l'extraction sémantique pour un scénario où zéro données parallèles est disponible.

- L'approche proposée est basée sur:
  (1) Un modèle extérieur permettant de générer une sequence de representations vectorielles à partir du texte
  (2) Un modèle SLU
  (3) Un modèle final



- Les résultats sont prometteuses et dépassent un simple modèle en cascade et les approches en utilisant de la voix synthétique

Merci pour votre attention

Salima Mdhaffar

https://sites.google.com/site/salimamedhaffar

- Post-doc au LIA
- Domaines de recherche:
  - Reconnaissance automatique de la parole
  - Compréhension de la parole
  - Apprentissage auto-supervisée
  - Apprentissage fédéré
- Projets actuels : Européen SELMA, ANR DeepPrivacy, LeBenchmark