XXXIVe Journées d'Études sur la Parole – JEP 2022
13-17 June 2022, Île de Noirmoutier, France

Architectures neuronales bout-en-bout pour la compréhension de la parole

Valentin Pelloin[1] Nathalie Camelin[1] Antoine Laurent[1]
Renato De Mori[2, 3] Sylvain Meignier[1]

[1] LIUM, Le Mans Université, 72085 Le Mans, France
[2] LIA, Avignon Université, 84029 Avignon, France
[3] Université de McGill, Montréal H3A 0G4, Canada
{prénom}.{nom}@univ-lemans.fr, rdemori@cs.mcgill.ca

RÉSUMÉ

Dans cet article, nous nous intéressons au problème de la compréhension de la parole et à sa résolution dans le cadre d'architectures dites *bout-en-bout*. Les différentes architectures proposées, basées sur des modèles neuronaux encodeurs-décodeurs avec mécanisme d'attention permettent d'émettre des hypothèses de contenus sémantiques directement à partir des caractéristiques acoustiques. Une première architecture a été conçue afin d'extraire à la fois les mots prononcés et les concepts. Testée sur le corpus MEDIA, elle permet une réduction d'erreur en absolu de 2,8 points par rapport à l'état de l'art. Avec cette même architecture, nous proposons une configuration originale permettant d'émettre également des hypothèses sur les valeurs des concepts. Enfin, une architecture composée de plusieurs décodeurs neuronaux chaînés pour un seul encodeur est testée dans l'objectif d'enrichir le décodeur d'informations linguistiques en plus des informations acoustiques.

ABSTRACT

End-to-end neural architectures for spoken language understanding

In this paper, we focus on *end-to-end* architectures designed to tackle spoken language understanding problems. We propose encoder-decoder architectures with an attention mechanism that focuses on relevant contextual acoustic features to hypothesize semantic contents. A first architecture has been built in order to extract pronounced words and concepts from speech. Tested on the MEDIA dataset, it obtains good results, and combined with a language model, it lowers the error by 2.8 points from the state-of-the-art results with end-to-end systems. With this same architecture, we propose a new configuration allowing to predict both concepts and their values. Lastly, a new architecture is proposed, composed of multiple chained decoders for a single encoder. With this architecture, we aim to improve the decoder with both linguistic and acoustic informations.

MOTS-CLÉS : compréhension de la parole, réseaux de neurones, mécanisme d'attention, séquence vers séquence, transfert d'apprentissage.

KEYWORDS: spoken language understanding, neural networks, attention mechanisms, sequence-to-sequence, transfer learning.

Ce travail est financé par le projet AISSPER, supporté par l'Agence Nationale de la Recherche (ANR), sous le contrat ANR-19-CE23-0004-01.

10.21437/JEP.2022-87

1 Introduction

Les systèmes de compréhension de la parole (Spoken Language Understanding - SLU) extraient les informations sémantiques contenues dans les tours de paroles issus de dialogues de diverses applications conversationnelles. Ces informations sémantiques sont généralement des fragments de l'ontologie du domaine d'application qui peut être définie selon un langage riche (*frame*) comme décrit par Tur & De Mori (2011). On réduit alors souvent le SLU en l'extraction de ces fragments (tâche de *slot filling*), appelés *concept* ainsi que leur *valeur* associée. Ainsi, dans une ontologie d'un système de dialogue de réservations touristiques, on pourra par exemple avoir une *frame ADRESSE* à laquelle appartient, entre autres, le *concept ville* dont une des valeurs peut être *Le Mans*. Dans les tours de parole, les concepts sont véhiculés par un ou plusieurs mots. Cette séquence de mots est notée *support*.

La compréhension peut être réalisée en deux temps, dans un système dit *cascade* ou alors en un temps, dans un système dit de *bout-en-bout*. Dernièrement, nous avons proposé dans Pelloin *et al.* (2021) un système neuronal de bout-en-bout qui permet d'extraire directement l'ensemble des couples concept/valeur depuis les caractéristiques acoustiques d'un tour de parole. Dans cette architecture originale, nous utilisons le mécanisme d'attention pour sélectionner efficacement l'ensemble des caractéristiques acoustiques nécessaires à la détection du concept et de sa valeur. Nous proposons ici une nouvelle architecture, basée sur cette première, qui intègre plusieurs décodeurs. Chaque décodeur est alors expert pour un type de sortie : transcription des mots, détection des concepts et des supports, détection des concepts et de leurs valeurs. En les chaînant efficacement, nous espérons améliorer les performances en utilisant en plus des caractéristiques acoustiques, les caractéristiques sémantiques et linguistiques grâce aux hypothèses des décodeurs précédents.

Le papier est organisé comme suit : la Section 2 présente brièvement l'état de l'art récent sur les systèmes de compréhension de la parole ainsi que des travaux sur des architectures neuronales utilisées dans d'autres domaines. Nos propres architectures sont décrites en Section 3 et évaluées en Section 4. Enfin, nous terminons par une conclusion et de nombreuses perspectives.

2 État de l'art

Ces dernières années, des architectures basées sur les réseaux de neurones profonds ont été proposées pour générer les hypothèses du domaine sémantique. Les premières architectures se composaient de deux systèmes utilisés en cascade : un système de transcription automatique de la parole (ASR) puis un système d'extraction de connaissances sémantiques (NLU) (Hakkani-Tür *et al.*, 2016; Zhang & Wang, 2016; Liu & Lane, 2016; Simonnet *et al.*, 2017). Par la suite, de nombreux travaux ont proposé d'intégrer ces modèles au sein d'une unique architecture, dite bout-en-bout (Qian *et al.*, 2017; Serdyuk *et al.*, 2018; Price *et al.*, 2020; Haghani *et al.*, 2018; Tomashenko *et al.*, 2020; Wang *et al.*, 2020).

Dans ce papier, une première architecture bout-en-bout est brièvement décrite. Celle-ci propose deux configurations. La première permet d'obtenir à la fois les mots prononcés et les concepts. Un système à base de règles, créé par un humain expert du domaine, est ensuite appliqué afin d'obtenir les valeurs. La seconde configuration permet d'obtenir directement à la fois les concepts et les valeurs. Précédemment présentée et testée par Pelloin *et al.* (2021), la première configuration obtenait des résultats état de l'art sur la tâche de la compréhension de la parole sur le corpus MEDIA (décrit en Section 4.1).

Souhaitant enrichir notre architecture d'informations linguistiques, en plus des informations acoustiques, nous proposons dans cet article une nouvelle architecture bout-en-bout (décrite en Section 3) composée de plusieurs décodeurs pour un seul encodeur. L'utilisation de multiples décodeurs au sein d'un même modèle a déjà été réalisée avec succès pour des tâches de traduction automatique. La compréhension de la parole dans notre cadre peut-être vue comme un problème de traduction automatique où le vocabulaire du langage cible serait l'ensemble des étiquettes de concepts et des valeurs possibles.

Anastasopoulos & Chiang (2018) utilisent une architecture dite *triangle*, pour la traduction, composée d'un encodeur et de deux décodeurs, où les sorties du premier décodeur (ASR) sont utilisées en entrée du second (traduction), avec celles de l'encodeur. Nous adaptons cette idée à la tâche de compréhension. En effet, nous pensons que l'extraction à la fois des concepts mais surtout des valeurs, qui elles sont basées uniquement sur le support de mots du concept, bénéficiera grandement des informations linguistiques du décodeur précédent en plus des informations acoustiques de l'encodeur. Toujours en traduction, un autre travail intéressant est initié par Garcia-Martinez *et al.* (2016), qui utilisent un seul décodeur augmenté, permettant de déterminer deux séquences de sorties différentes simultanément.

Par ailleurs, les récentes avancées des modèles neuronaux dits *transformers*, et plus particulièrement des modèles autosupervisés dérivés de l'architecture BERT (Devlin *et al.*, 2019) ont permis une avancée considérable dans de nombreux domaines du traitement automatique de la langue (Devlin *et al.*, 2019; Martin *et al.*, 2020; Baevski *et al.*, 2020). Appliqués à la tâche de compréhension de la parole, ces modèles restent néanmoins utilisés en suivant une approche cascade. Ghannay *et al.* (2021) utilisent d'abord un premier modèle wav2vec 2.0 (Baevski *et al.*, 2020) afin d'extraire la séquence de mots la plus probable, puis utilisent un modèle CamemBERT (Martin *et al.*, 2020) affiné sur la tâche NLU. Les résultats obtenus avec cet ensemble définissent le nouvel état de l'art de la tâche MEDIA. Ils sont obtenus avec des systèmes en cascade et utilisent un système à base de règles expert pour obtenir les valeurs.

3 Architectures SLU bout-en-bout

Notre modèle de base est celui présenté par Pelloin *et al.* (2021). Il s'agit d'un réseau de neurones encodeur-décodeur avec mécanisme d'attention. Nous proposons dans ce papier, une nouvelle version basée sur notre architecture de base mais enrichie afin que le décodeur pour la compréhension puisse bénéficier à la fois des caractéristiques acoustiques, mais aussi linguistiques.

3.1 Système simple décodeur

Architecture. Notre architecture neuronale encodeur-décodeur construite avec un mécanisme d'attention est basée sur une recette *Espresso* (Wang *et al.*, 2019), initialement conçue pour la tâche de transcription automatique de la parole du corpus *Wall Street Journal*.

L'entrée du modèle est constituée de 40 filtres Mel (*MelFBanks*) extraits avec une fenêtre de Hamming de 25ms et 10ms de pas. La sortie est une séquence de caractères permettant l'obtention des mots, des valeurs normalisées et/ou des étiquettes de concepts.

Soit $X = (x_1, ..., x_{T_x})$ le vecteur de paramètres d'entrée du modèle et $Y = (y_1, ..., y_{T_y})$ les sorties

du modèle. Notre modèle détermine automatiquement la séquence de sortie Y à partir de la séquence d'entrée X. L'encodeur est tout d'abord composé de 4 blocs convolutionels 2D, où chaque couche de convolution est suivie d'une normalisation par lots. Ensuite, 4 couches récurrentes bidirectionnelles de LSTM (biLSTM) sont superposées afin de générer la représentation cachée de l'encodeur :

$$\text{Encodeur}(X) = (\overleftrightarrow{h_1}, ..., \overleftrightarrow{h_{T_x}}) \tag{1}$$

Le décodeur utilise 4 LSTM suivies de 2 couches linéaires et une fonction softmax. Les états cachés du décodeur sont calculés en utilisant un mécanisme d'attention, qui aligne les états cachés de l'encodeur et les précédents états cachés du décodeur :

$$Y = \text{Décodeur}(H^{\text{att}}) \tag{2}$$

Le mécanisme d'attention H^{att} appliqué est décrit par Bahdanau *et al.* (2015).

Apprentissage. Notre procédure d'apprentissage s'inspire de celle de transfert d'apprentissage par curriculum proposée par Caubrière *et al.* (2019). Nous apprenons tout d'abord un modèle pour la première tâche de transcription automatique à l'aide de corpus hors-domaine et du domaine, tel que décrit dans la section 4.1. Ensuite, nous affinons ce modèle en utilisant uniquement les corpus du domaine pour obtenir un système de compréhension.

3.2 Architecture multi-décodeurs

Le modèle encodeur-décodeur avec mécanisme d'attention présenté dans la Section précédente est transformé en un modèle comportant un encodeur pour de multiples décodeurs. Chaque décodeur possède ses propres mécanismes d'attention. Ces décodeurs sont chaînés entre eux : les informations déterminées par le premier décodeur sont transmises en entrée des décodeurs suivants, en plus de celles issues de l'encodeur. De cette manière, l'hypothèse en sortie du décodeur suivant est basée sur des sources d'informations diverses.

L'architecture proposée est présentée dans la Figure 1 pour un exemple à deux décodeurs. L'encodeur génère une représentation du signal de parole. Un premier décodeur, nommé *ASR* doit réaliser la transcription automatique de ce signal de parole. Un second décodeur, nommé *SLU$_{Mots}$*, doit lui réaliser la tâche d'extraction des mots et des concepts, en travaillant sur les deux modalités existantes : la représentation du signal de parole provenant de l'encodeur, ainsi que la représentation de la transcription automatique de ce signal, provenant du décodeur *ASR*.

Dans un tel modèle, trois mécanismes d'attention sont utilisés :

— au sein du décodeur ASR, un mécanisme d'attention permet l'alignement entre la représentation provenant de l'encodeur et les caractères ASR de sortie ;

— au sein du décodeur SLU$_{\text{Mots}}$, un premier mécanisme d'attention permet également cet alignement entre la représentation de l'encodeur et les caractères SLU de sortie, et un second permet l'alignement entre les caractères ASR du décodeur ASR et les caractères SLU. Pour chaque caractère de sortie SLU, les deux vecteurs de contexte des mécanismes d'attention sont concaténés afin de générer une représentation commune.

Cette architecture multi-décodeurs peut être étendue, et le nombre de décodeurs chaînés n'est pas limité à deux. Dans une architecture simple décodeur, seul le décodeur ASR est conservé et optimisé. Dans une architecture comportant trois décodeurs, un troisième décodeur est ajouté au second de

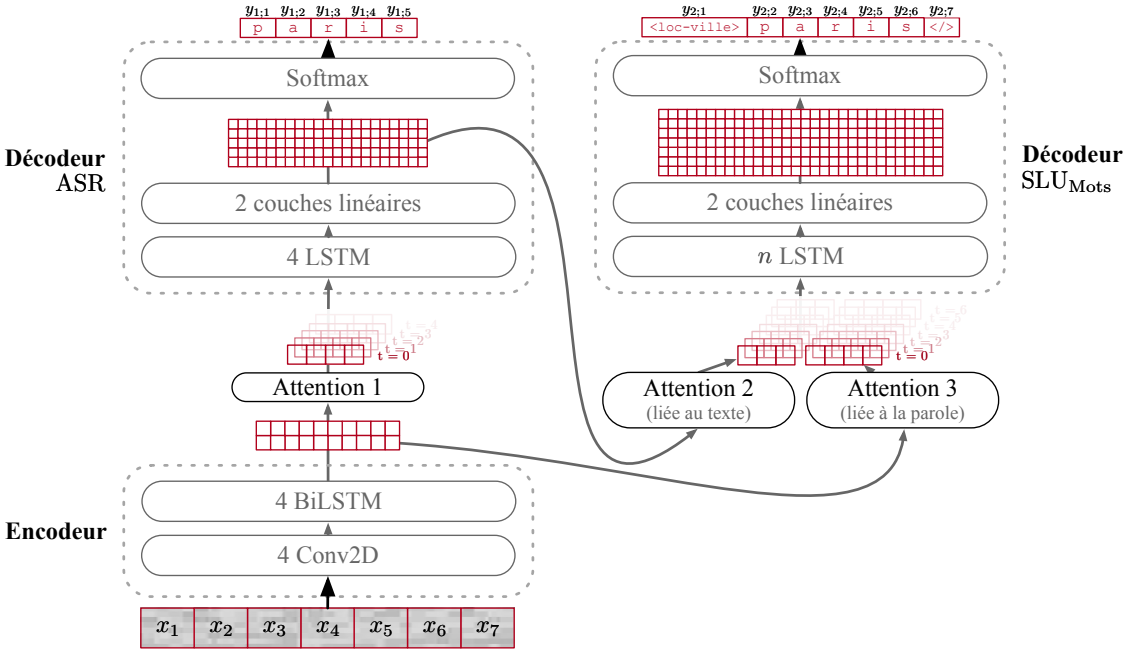

FIGURE 1 – Architecture d'un modèle multi-décodeurs comportant deux décodeurs chaînés : un décodeur ASR et un décodeur SLU$_{\text{Mots}}$. Le nombre n de couches LSTM dans le deuxième décodeur est ajusté en fonction de la tâche à réaliser pour celui-ci.

la même manière que le second au premier pour réaliser par exemple directement la tâche finale d'extraction de concepts et de valeurs.

4 Expériences

4.1 Jeux de données

Le modèle de transcription (décodeur ASR) est appris en utilisant les différents corpus français d'actualité décrits par Caubrière *et al.* (2019) et également des corpus de conversations téléphoniques du domaine : PORTMEDIA (Lefèvre *et al.*, 2012), MEDIA (Devillers *et al.*, 2004) et DECODA (Bechet *et al.*, 2012). Un total de 414 heures d'apprentissage est ainsi utilisé pour l'ASR.

Un tel modèle requiert une large quantité de données audio, ainsi que la transcription manuelle de celle-ci. Cependant, l'alignement des trames acoustiques avec les mots n'est pas nécessaire, notre système apprendra lui-même l'alignement entre les caractères de sortie et les trames d'entrée, grâce à son mécanisme d'attention (Chorowski *et al.*, 2014; Chan *et al.*, 2016; Bahdanau *et al.*, 2016).

Le modèle de compréhension *SLU$_{Mots}$* est lui appris seulement sur les données du domaine très proches de la tâche, *c.à.d* les 24 heures d'apprentissage de MEDIA et PORTMEDIA. Ces deux corpus sont constitués de dialogues téléphoniques entre deux personnes, dans un mode dit de *magicien d'Oz* où une personne joue le rôle d'un ordinateur tandis que l'autre joue le rôle d'un utilisateur souhaitant obtenir des informations. Le corpus MEDIA comporte des requêtes de réservation et d'informations

sur des hôtels, et PORTMEDIA comporte des conversations à propos de pièces de théâtre durant le Festival d'Avignon. Toutes les conversations ont été transcrites manuellement, et les tours de paroles de l'utilisateur sont annotés avec des concepts sémantiques, chaque support est délimité et la valeur est normalisée. Un exemple est donné dans la Table 1 : le concept *lienref-coref* est véhiculé par le support *l'*, normalisé par la valeur *singulier*.

Exemple annoté	{est ce qu' il y a ; - ; -} {une piscine ; **hôtel-services** ; *piscine*} {à ; - ; -}
{*mots ; concept ; valeur*}	{l' ; **lienref-coref** ; *singulier*} {hôtel ; **objet-bd** ; *hotel*}
ASR	est ce qu' il y a une piscine à l' hôtel
SLU$_{\text{Mots}}$	est ce qu' il y a **hôtel-services** une piscine </> à **lienref-coref** l' </> **objet-bd** hôtel </>
SLU$_{\text{Norm}}$	**hôtel-services** piscine </> **lienref-coref** singulier </> **objet-bd** hotel </>

TABLE 1 – Un exemple annoté du corpus MEDIA et ses différentes représentations utilisées en sortie des décodeurs.

Béchet & Raymond (2019) ont montré que MEDIA était un des corpus les plus difficiles dans le domaine de la compréhension de la parole. En effet, il est composé de nombreux concepts directement liés au domaine applicatif (*e.g. hôtel-services*) mais aussi des concepts sémantiques généraux comme des connecteurs logiques basés sur des supports comme "*et*" ou "*ou*" qui se réfèrent à des concepts mentionnés auparavant. D'autres concepts sont également très compliqués à extraire automatiquement comme les co-références qui nécessitent la connaissance de plusieurs tours de parole afin d'être résolus.

Les résultats présentés dans ce papier portent uniquement sur le corpus MEDIA, découpés en respectant la segmentation officielle du corpus d'apprentissage, développement, et test. MEDIA contient respectivement 727, 79 et 208 dialogues, contenant 12,9k, 1,3k et 3,5k tours de paroles de l'utilisateur. Ces tours de paroles sont annotés en utilisant un lexique sémantique de 76 concepts, qui occurrent respectivement dans 31,7k, 3,3k et 8,8k supports.

4.2 Apprentissage et décodage

Dans le cas du modèle simple décodeur, un premier apprentissage de l'architecture est effectué sur les données ASR pour obtenir un système de transcription. L'apprentissage est ensuite repris sur les données SLU MEDIA et PORTMEDIA pour obtenir un système de compréhension de la parole. La transition entre ASR et SLU$_{\text{Mots}}$/SLU$_{\text{Norm}}$ est effectuée en augmentant la sortie avec les symboles des 76 concepts.

En ce qui concerne les modèles comportant plusieurs décodeurs, nous augmentons le modèle de base en créant les décodeurs désirés, selon les sorties voulues. Le décodeur ASR est conservé et continue à être optimisé lors de l'ajout des décodeurs suivants. La fonction de coût est calculée en faisant la somme pondérée des fonctions de coûts de chacun des décodeurs.

Toutes les sorties possibles des décodeurs sont présentées en Table 1 sur un exemple du corpus MEDIA. Pour des raisons de clarté, les séquences correspondant aux mots sont représentées à l'aide de mots, mais les modèles doivent générer un à un chacun des caractères. Les concepts correspondent eux à une seule unité de sortie.

Un décodeur de type SLU$_{\text{Norm}}$ est introduit afin d'émettre directement une sortie contenant les concepts et leurs valeurs associées. Dans cette variante, les caractères de sortie du modèle représentent soit les

concepts soit les caractères composant les valeurs normalisées (et non les caractères correspondant aux mots prononcés). Lorsque le décodeur SLU_{Mots} est utilisé, l'extraction des valeurs normalisées est réalisée à l'aide de règles déterminées manuellement par un expert humain. Ces règles, utilisées dans de nombreux travaux sur MEDIA (Tomashenko *et al.*, 2020; Caubrière *et al.*, 2019; Ghannay *et al.*, 2018; Simonnet *et al.*, 2017), s'appliquent sur les concepts et leurs supports. Il va de soi que le système SLU_{Mots} est avantagé par rapport à SLU_{Norm} car il utilise de la connaissance humaine pour obtenir les valeurs. Néanmoins le système SLU_{Norm} a lui l'avantage de nous permettre de traiter en totalité le problème de la compréhension de manière automatique.

4.3 Résultats

Les sorties sont évaluées selon les métriques du *Concept Error Rate* (CER) et du *Concept-Value Error Rate* (CVER). L'alignement entre les séquences de référence et celles hypothèses émises par les décodeurs produit un nombre d'erreurs d'insertions, de substitutions et de suppressions, qui sont utilisées, de manière identique au WER (*Word Error Rate*), pour calculer l'erreur totale du modèle.

À titre de comparaison, la Table 2 présente les résultats obtenus avec le modèle simple décodeur optimisé comme décrit par Pelloin *et al.* (2021) : la méthode par transfert d'apprentissage a été appliquée jusqu'au corpus MEDIA uniquement, les résultats sont obtenus à l'aide d'un décodage par recherche en faisceau (*beam search*) et un modèle de langage neuronal additionnel, appris sur les données du domaine est utilisé.

	Décodeurs	**décodage en *beam***	
		%CER	%CVER
(a)	SLU_{Mots}	13,6	18,5
(b)	SLU_{Norm}	15,4	21,6

TABLE 2 – Résultats sur le corpus MEDIA Test du modèle *optimisé* simple décodeur.

Ces résultats nous ont permis de montrer l'intérêt du mécanisme d'attention (en obtenant le nouvel état de l'art sur ce corpus et un gain de 2,8% en CER par rapport à (Caubrière *et al.*, 2019)). De plus, nous étions les premiers à proposer une chaîne complètement automatique pour l'obtention des concepts et des valeurs avec le décodeur SLU_{Norm} à partir des caractéristiques acoustiques. Cette architecture performante nous a servi de base pour notre nouvelle proposition dont les premiers résultats sont présentés dans la Table 3.

#		**Décodeurs**			**décodage en *beam***	
					%CER	**%CVER**
1	(c)	SLU_{Mots}			16,74	21,51
2	(d)	ASR	SLU_{Mots}		16,58	22,70
	(e)	ASR	SLU_{Norm}		18,98	26,00
3	(f)	ASR	SLU_{Mots}	SLU_{Norm}	16,40/22,28*	22,31/27,77*

TABLE 3 – Premiers résultats sur le corpus MEDIA Test des différentes configurations explorées de l'architecture multi-décodeurs. *Évaluation de SLU_{Mots}/SLU_{Norm}.

Les résultats présentés ici sont ceux d'un système pas encore totalement optimisé : pas d'ajout de modèles de langage ni d'optimisation des modèles sur le corpus MEDIA seul. Il s'agit, dans un

premier temps, de valider notre proposition d'architecture multi-décodeurs. Les optimisations qui ont montré leur intérêt sur le modèle simple décodeur seront apportées dans nos futurs travaux.

À titre de comparaison, la ligne 1 présente les résultats du modèle simple décodeur sans optimisation. On remarque que les différentes optimisations proposées par Pelloin *et al.* (2021) nous permettent un gain absolu de 3 points en passant de 16,74% à 13,6% de CER.

Dans le cas où le décodeur est SLU_{Mots}, nous remarquons qu'il y a effectivement un gain lorsque l'on passe à l'architecture deux décodeurs (système (c) comparé avec (d)) avec 16,58 de CER et même 16,40 (système ligne (f)) avec l'architecture à trois décodeurs. La tendance d'amélioration n'est pas observée si on considère également les valeurs. Avec une perte de 21,51 à 22,70 pour l'architecture deux décodeurs et un CVER à 22,31 pour celle à trois décodeurs.

Concernant les résultats du décodeur SLU_{Norm}, nous pouvons simplement remarquer une dégradation du CVER de seulement 3 points pour le système (e) par rapport au système (d) qui, rappelons-le, utilise des règles humaines pour obtenir les valeurs. Trois points de pertes étaient également observés dans la version simple décodeur optimisé (systèmes (a) et (b)). En revanche, dans cette configuration, l'architecture à trois décodeurs n'est pour l'instant pas prometteuse. Proposant une architecture plus complexe, elle nécessite certainement des ajustements au niveau de son optimisation.

5 Conclusions et perspectives

Dans cet article, nous proposons plusieurs modèles de compréhension de la parole basés sur une architecture, dite bout-en-bout, de type encodeur-décodeurs utilisant des mécanismes d'attention et des décodeurs neuronaux travaillant conjointement. Le premier modèle qui se compose d'un simple encodeur-décodeur, optimisé pour réaliser la tâche de transcription puis de compréhension, a établi l'état de l'art sur le corpus MEDIA dans Pelloin *et al.* (2021). Nous proposons ici une extension de ce modèle à plusieurs décodeurs afin d'améliorer notamment les résultats de notre configuration SLU_{Norm} qui permet d'obtenir de manière complètement automatique l'ensemble des concepts/valeurs nécessaires à la réalisation de la tâche de compréhension sur le corpus MEDIA. Les premiers résultats sont encourageants en ce qui concerne l'architecture double décodeurs. Nous souhaitons par la suite optimiser cette architecture comme proposé dans Pelloin *et al.* (2021) pour améliorer les performances (intégration de modèles de langage, amélioration du décodage de type *beam search*, optimisation sur MEDIA).

Par ailleurs, un nouvel état de l'art a été établi sur MEDIA par Ghannay *et al.* (2021). Leur modèle est de type cascade et utilise notamment BERT dans le module de compréhension. Avec notre nouvelle architecture double décodeurs, nous pensons pouvoir intégrer l'information de ces modèles entre la sortie du décodeur ASR et l'entrée du décodeur SLU. Nous espérons ainsi profiter de ces larges modèles pré-entraînés sur des quantités importantes de données tout en restant dans une architecture bout-en-bout qui ne perd pas d'informations (acoustiques notamment) à cause d'une représentation intermédiaire. Nous pourrions également intégrer à ce niveau des informations syntaxiques (comme les étiquettes morphosyntaxiques), sémantiques comme cela était le cas dans Simonnet *et al.* (2017) ou encore des informations sur l'historique du dialogue.

Références

ANASTASOPOULOS A. & CHIANG D. (2018). Tied multitask learning for neural speech translation. In *NAACL 2018*, p. 82–91, New Orleans, Louisiana.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *NeurIPS 2020*, **33**, 12449–12460.

BAHDANAU D., CHO K. & BENGIO Y. (2015). Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

BAHDANAU D., CHOROWSKI J., SERDYUK D., BRAKEL P. & BENGIO Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *ICASSP 2016*, p. 4945–4949.

BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BEZE M., MORI R. D. & ARBILLOT E. (2012). Decoda : a call-centre human-human spoken conversation corpus. In *LREC 2012*, p. 1343–1347.

BÉCHET F. & RAYMOND C. (2019). Benchmarking benchmarks : introducing new automatic indicators for benchmarking Spoken Language Understanding corpora. In *Interspeech 2019*, Graz, Austria.

CAUBRIÈRE A., TOMASHENKO N., LAURENT A., MORIN E., CAMELIN N. & ESTÈVE Y. (2019). Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability. *Interspeech 2019*, p. 1198–1202.

CHAN W., JAITLY N., LE Q. V. & VINYALS O. (2016). Listen, attend and spell : A neural network for large vocabulary conversational speech recognition. In *ICASSP 2016*.

CHOROWSKI J., BAHDANAU D., CHO K. & BENGIO Y. (2014). End-to-end continuous speech recognition using attention-based recurrent nn : First results. In *NIPS 2014*.

DEVILLERS L., MAYNARD H., ROSSET S., PAROUBEK P., MCTAIT K., MOSTEFA D., CHOUKRI K., CHARNAY L., BOUSQUET C., VIGOUROUX N., BÉCHET F., ROMARY L., ANTOINE J. Y., VILLANEAU J., VERGNES M. & GOULIAN J. (2004). The French MEDIA/EVALDA project : the evaluation of the understanding capability of spoken language dialogue systems. In *LREC 2004*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *NAACL 2019*, p. 4171–4186, Minneapolis, Minnesota.

GARCIA-MARTINEZ M., BARRAULT L. & BOUGARES F. (2016). Factored Neural Machine Translation Architectures. In *IWSLT 2016*, Seattle, United States.

GHANNAY S., CAUBRIÈRE A., ESTÈVE Y., CAMELIN N., SIMONNET E., LAURENT A. & MORIN E. (2018). End-to-end named entity and semantic concept extraction from speech. In *SLT 2018*, Athens, Greece.

GHANNAY S., CAUBRIÈRE A., MDHAFFAR S., LAPERRIÈRE G., JABAIAN B. & ESTÈVE Y. (2021). Where Are We in Semantic Concept Extraction for Spoken Language Understanding ? *SPECOM 2021*, p. 202–213.

HAGHANI P., NARAYANAN A., BACCHIANI M., CHUANG G., GAUR N., MORENO P., PRABHAVALKAR R., QU Z. & WATERS A. (2018). From Audio to Semantics : Approaches to End-to-End Spoken Language Understanding. *SLT 2018*, p. 720–726.

HAKKANI-TÜR D., TÜR G., CELIKYILMAZ A., CHEN Y. N., GAO J., DENG L. & WANG Y. Y. (2016). Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Interspeech 2016*, p. 715–719, San Francisco, USA.

LEFÈVRE F., MOSTEFA D., BESACIER L., ESTÈVE Y., QUIGNARD M., CAMELIN N., FAVRE B., JABAIAN B. & ROJAS B. L. M. (2012). Leveraging study of robustness and portability of spoken language understanding systems across languages and domains : the PORTMEDIA corpora. In *LREC 2012*.

LIU B. & LANE I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. In *Interspeech 2016*, p. 685–689.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a Tasty French Language Model. In *ACL 2020*, p. 7203–7219, Stroudsburg, PA, USA.

PELLOIN V., CAMELIN N., LAURENT A., DE MORI R., CAUBRIÈRE A., ESTÈVE Y. & MEIGNIER S. (2021). End2End Acoustic to Semantic Transduction. In *ICASSP 2021*, Toronto, ON, Canada.

PRICE R., MEHRABANI M. & BANGALORE S. (2020). Improved End-To-End Spoken Utterance Classification with a Self-Attention Acoustic Classifier. *ICASSP 2020*, p. 8504–8508.

QIAN Y., UBALE R., RAMANARYANAN V., LANGE P., SUENDERMANN-OEFT D., EVANINI K. & TSUPRUN E. (2017). Exploring ASR-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system. *ASRU 2017*, p. 569–576.

SERDYUK D., WANG Y., FUEGEN C., KUMAR A., LIU B. & BENGIO Y. (2018). Towards End-to-end Spoken Language Understanding. *ICASSP 2018*, p. 5754–5758.

SIMONNET E., GHANNAY S., CAMELIN N., ESTÈVE Y. & DE MORI R. (2017). ASR error management for improving spoken language understanding. In *Interspeech 2017*.

TOMASHENKO N., RAYMOND C., CAUBRIERE A., DE MORI R. & ESTÈVE Y. (2020). Dialogue History Integration into End-to-End Signal-to-Concept Spoken Language Understanding Systems. *ICASSP 2020*, p. 8509–8513.

TUR G. & DE MORI R. (2011). Chapter 1 : Spoken language understanding for human/machine interactions. In *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*.

WANG P., WEI L., CAO Y., XIE J. & NIE Z. (2020). Large-scale unsupervised pre-training for end-to-end spoken language understanding. *ICASSP 2020*, p. 7994–7998.

WANG Y., CHEN T., XU H., DING S., LV H., SHAO Y., PENG N., XIE L., WATANABE S. & KHUDANPUR S. (2019). Espresso : A fast end-to-end neural speech recognition toolkit. In *ASRU 2019*, p. 136–143, Singapore, Singapore.

ZHANG X. & WANG H. (2016). A joint model of intent determination and slot filling for spoken language understanding. In *IJCAI 2016*, p. 2993–2999.