Curriculum d'apprentissage : reconnaissance d'entités nommées pour l'extraction de concepts sémantiques

Antoine Caubrière[1] Natalia Tomashenko[2] Yannick Estève[2]
Antoine Laurent[1] Emmanuel Morin[3]
(1) LIUM, Avenue Olivier Messiaen, 72085 Le Mans, France
(2) LIA, 339 Chemin des Meinajaries, 84140 Avignon, France
(3) LS2N, 2 Chemin de la Houssinière, 44322 Nantes, France
prénom.nom@{univ-lemans,univ-avignon,univ-nantes}.fr

RÉSUMÉ

Dans cet article, nous présentons une approche de bout en bout d'extraction de concepts sémantiques de la parole. En particulier, nous mettons en avant l'apport d'une chaîne d'apprentissage successif pilotée par une stratégie de curriculum d'apprentissage. Dans la chaîne d'apprentissage mise en place, nous exploitons des données françaises annotées en entités nommées que nous supposons être des concepts plus génériques que les concepts sémantiques liés à une application informatique spécifique. Dans cette étude, il s'agit d'extraire des concepts sémantiques dans le cadre de la tâche MEDIA. Pour renforcer le système proposé, nous exploitons aussi des stratégies d'augmentation de données, un modèle de langage 5-gramme, ainsi qu'un mode étoile aidant le système à se concentrer sur les concepts et leurs valeurs lors de l'apprentissage. Les résultats montrent un intérêt à l'utilisation des données d'entités nommées, permettant un gain relatif allant jusqu'à 6,5 %.

ABSTRACT

Curriculum learning : named entity recognition for semantic concept extraction

In this paper, we present an end-to-end approach for semantic concept extraction from speech. In particular, we highlight the contribution of a successive learning chain driven by a curriculum learning strategy. In the learning chain, we use French data with named entity annotations that we assume are more generic concepts than semantic concept related to a specific computer application. In this study, the aim is to extract semantic concept as part of the MEDIA task. To improve the proposed system, we also use data augmentation, 5-gram langage model and a star mode to help the system focus on concepts and their values during the training. Results show an interest in using named entity data, allowing a relative gain up to 6.5%.

MOTS-CLÉS : Curriculum d'apprentissage, transfert d'apprentissage, bout en bout, extraction de concepts sémantiques, entités nommées.

KEYWORDS: Curriculum learning, transfer learning, end-to-end, semantic concept extraction, named entity.

1 Introduction

L'apprentissage humain est réalisé par étapes successives de plus en plus complexes, permettant ainsi d'aborder des notions ordonnées de la plus simple à la plus compliquée. Basés sur cette observation, des travaux ont été menés afin d'appliquer ce concept aux algorithmes d'apprentissage automatique (Bengio *et al.*, 2009). Ces travaux ont montré l'intérêt de l'organisation des exemples d'apprentissage d'un même ensemble de données pour l'amélioration de la vitesse de convergence et des performances de généralisation des systèmes entraînés. Il s'agit du curriculum d'apprentissage.

D'autres travaux se sont concentrés sur l'application des principes de transfert d'apprentissage (Weinshall *et al.*, 2018; Pan & Yang, 2010) pour ordonner les exemples appris par le système. Ils ont de nouveau montré l'intérêt d'un curriculum d'apprentissage, notamment pour l'amélioration des performances de généralisation pour des tâches complexes.

Récemment, il a été proposé par Ghannay *et al.* (2018) une première approche de bout en bout permettant de réaliser, de manière totalement disjointe, soit une reconnaissance d'entités nommées, soit une extraction de concepts sémantiques dans la parole. La proposition consistait en l'ajout des frontières des concepts directement dans les séquences à produire par le système. Traditionnellement, les tâches d'extraction de concepts sémantiques et de reconnaissance d'entités nommées dans la parole s'effectuent par l'intermédiaire d'une chaîne de composants. Un système de reconnaissance de la parole constitue le premier composant, puis un système de traitement du langage est appliqué sur les transcriptions automatiques produites par ce premier composant. Le système de traitement du langage est appliqué sur des transcriptions imparfaites. L'avantage d'un système de bout en bout réside dans sa possibilité à limiter la transmission des erreurs de transcription de la parole. Mais aussi, dans l'optimisation jointe des composants de parole et de traitement de la langue pour la tâche finale.

Dans cet article, nous présentons un système qui s'appuie entièrement sur une architecture neuronale similaire au système de reconnaissance de la parole DeepSpeech 2 de Baidu (Amodei *et al.*, 2016). Nous entraînons ce système pour réaliser l'extraction des concepts sémantiques dans les données MEDIA (Devillers *et al.*, 2004). En considérant que la reconnaissance d'entités nommées et l'extraction de concepts sémantiques (comme proposé dans la campagne d'évaluation MEDIA) sont deux tâches proches, nous souhaitons explorer la possibilité d'exploiter des données d'entités nommées pour améliorer notre système d'extraction de concepts. L'application d'un principe de curriculum d'apprentissage est motivée par l'observation du caractère plus générique des annotations d'entités nommées par rapport aux annotations de concepts sémantiques.

Dans nos travaux, nous adaptons le curriculum pour mettre en place une chaîne d'entraînements successifs ordonnés du plus général au plus spécifique. Il s'agit d'une chaîne d'apprentissage pilotée par une stratégie de curriculum. Elle exploite le transfert d'apprentissage, qui nous permet de faire face au manque de données annotées pour la tâche finale. Nous utilisons également un modèle de langage 5-gramme, ainsi qu'un mode étoile permettant de concentrer le système sur les concepts et leurs valeurs. Notre meilleur système obtient des performances bien meilleures que l'état de l'art, qu'il conviendra toutefois de nuancer.

La structure de cet article est la suivante : nous présentons tout d'abord des travaux similaires dans la deuxième section. Puis, dans la troisième section, nous expliquons l'architecture neuronale utilisée. Ensuite, la quatrième section est dédiée à la présentation des ensembles de données et du mode étoile que nous utilisons dans nos travaux. La section suivante présente la chaîne d'apprentissage mise en place. Enfin, l'avant dernière section est consacrée à nos résultats expérimentaux avant de conclure.

2 Travaux similaires

Les récentes avancées des systèmes de reconnaissance de la parole de bout en bout permettent désormais d'obtenir des performances solides (Amodei *et al.*, 2016). Des travaux ont été réalisés sur des tâches appliquées à la parole. Certains se concentrent sur la traduction automatique de bout en bout, faisant suite aux travaux précurseurs de Bérard *et al.* (2016). Ils mettent en place un système neuronal de type encodeur-décodeur, basé sur un mécanisme d'attention. Ce système est entraîné à l'aide d'un petit ensemble de données français parlé - anglais écrit et a la particularité d'utiliser de la parole de synthèse. Les travaux de Weiss *et al.* (2017) sont similaires mais utilisent des données réelles. Ils exploitent un modèle neuronal initialement utilisé en reconnaissance de la parole. Un système du même type est aussi exploité dans les travaux de Bérard *et al.* (2018), dont l'originalité tiens dans l'augmentation de données d'apprentissage par l'utilisation d'un système de traduction texte vers texte. Ces approches sont prometteuses, mais n'ont pas atteint des performances à l'état de l'art, comme le montrent les résultats de la campagne d'évaluation IWSLT 2018 pour la traduction de l'anglais parlé vers l'allemand écrit (Jan *et al.*, 2018).

D'autres travaux concernant la détection d'intention et la détection de domaine ont été réalisés par Serdyuk *et al.* (2018). Ces travaux mettent en œuvre un réseau de neurones inspiré des technologies de reconnaissance de la parole. Ce système obtient d'excellents résultats sur la tâche de détection de domaine, mais n'arrive pas à atteindre des performances à l'état de l'art pour la tâche de détection d'intention. Il montre toutefois l'intérêt de partir directement du signal de parole pour la tâche de compréhension du langage. Des conclusions similaires sont partagées par Ghannay *et al.* (2018) pour la reconnaissance des entités nommées ou l'extraction de concepts sémantiques dans le cadre d'une tâche de *slot filling*.

Enfin, les récents travaux de Platanios *et al.* (2019) ont montré l'intérêt d'une approche exploitant l'apprentissage par curriculum pour une tâche de traduction automatique. Dans ces travaux, les auteurs ont défini les notions de « difficulté » d'un exemple d'apprentissage et de « compétence » d'un modèle appris, dans le cadre de la traduction automatique. Ces deux notions permettent de filtrer les exemples d'apprentissage pour les présenter des plus simples au plus compliqués.

3 Système neuronal

L'architecture neuronale que nous utilisons pour ces travaux est très similaire au système de reconnaissance de la parole DeepSpeech 2 de Baidu (Amodei *et al.*, 2016). Il est composé d'un empilement de couches de convolution (CNN), de couches récurrentes bidirectionnelles (biLSTM), d'une couche de convolution lookahead, d'une couche entièrement connectée et enfin d'une couche softmax. Ce système utilise en entrée des log-spectrogrammes de l'audio calculés sur des fenêtres de 20ms et prédit des séquences de caractères.

Pour nos expérimentations, nous utilisons une implémentation [1] avec deux couches CNN et cinq couches biLSTM. Notre sortie peut correspondre soit à la meilleure hypothèse produite par le système, soit à une hypothèse recalculée à l'aide de l'algorithme « beam search » et d'un modèle de langage 5-gramme. L'architecture neuronale est synthétisée en figure 1.

1. https ://github.com/SeanNaren/deepspeech.pytorch

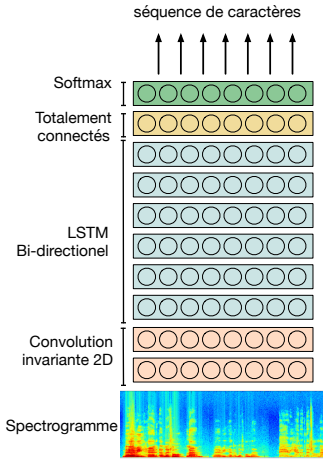

FIGURE 1 – Architecture neuronale DeepSpeech 2

Ce système est entraîné de bout en bout en utilisant la fonction de coût Connectionist Temporal Classification (CTC) (Graves *et al.*, 2006). Une particularité de cette fonction est qu'elle ne nécessite pas d'alignement a priori entre les entrées et les sorties du système. Il est appris pendant l'entraînement.

Pour apprendre cet alignement, les log-spectrogrammes des séquences d'audio sont découpés en tranches. Il s'agit de tranches de taille arbitraire fixe permettant la prédiction d'étiquettes de sorties. Pour chaque tranche, une distribution de probabilités est calculée sur l'ensemble des étiquettes prédictibles (caractères). La séquence produite est constituée de l'ensemble des étiquettes les plus probables de chaque tranche.

Comme la parole humaine est fluctuante en vitesse, un caractère est susceptible d'être représenté par plusieurs tranches. C'est pourquoi, les répétitions sont supprimées dans les séquences produites. Cette réduction empêche l'alignement correct d'un segment audio avec une séquence contenant normalement des répétitions (« curriculum » deviendrait « curiculum »). Pour prendre en compte les répétitions, un caractère spécifique est ajouté dans les étiquettes prédictibles par le système, noté ϵ dans l'exemple de la figure 2. Cette étiquette permet de délimiter deux caractères identiques successifs. Elle ne représente donc aucune information devant être conservée et sera supprimée des séquences finales produites après réduction des répétitions. De plus, le flux de parole n'est pas systématiquement constant sur un segment, il peut être composé de silence. Il est donc inutile de forcer l'alignement de ces portions à une étiquette porteuse de sens. Le caractère spécifique utilisé pour gérer les répétitions est aussi utilisé pour les silences composant l'audio. La figure 2 représente un exemple de l'alignement appris par la fonction CTC. La modélisation de séquence par l'intermédiaire de cette fonction est davantage détaillée par Hannun (2017).

L'alignement entre l'audio et les séquences à produire est appris pendant l'entraînement. Ainsi, l'information relative aux entités nommées et aux concepts sémantiques peut être injectée dans les séquences cibles. Dans nos travaux, nos modifications interviennent au niveau des séquences de caractères présentées au système. Les frontières des concepts y sont ajoutées. Le système produit donc des séquences de caractères au sein desquelles des caractères spécifiques représentent les frontières des concepts.

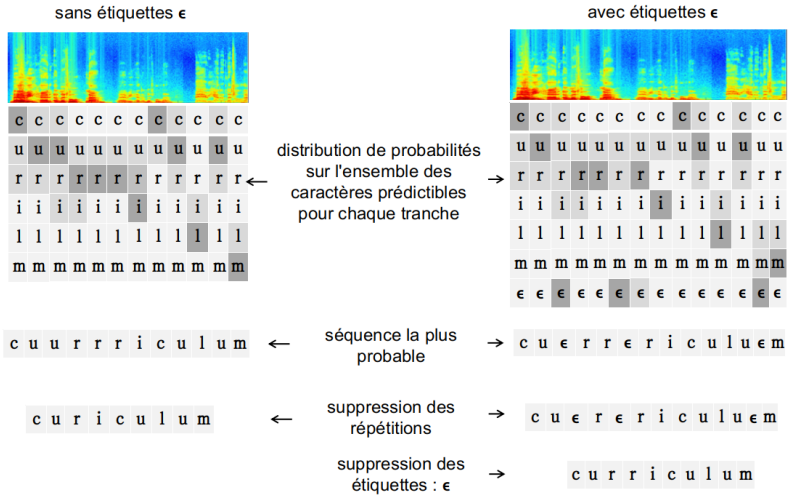

FIGURE 2 – exemple d'alignements de la fonction CTC

4 Données, Augmentation de données, Mode étoile

Dans ces travaux, nous utilisons plusieurs ensembles de données dans le but de maximiser notre quantité de données audio, de données annotées en entités nommées et de données annotées en concepts sémantiques.

4.1 Audio

Parmi les données audio que nous utilisons dans cette étude, une grande partie est issue d'enregistrements d'émissions de radios et de chaînes de télévision francophone. Chaque ensemble est transcrit manuellement et est séparé en trois parties distinctes : développement, test et apprentissage. Les corpus utilisés sont : EPAC (Esteve *et al.*, 2010), ESTER2 (Galliano *et al.*, 2009), ETAPE (Gravier *et al.*, 2012), QUAERO (Grouin *et al.*, 2011), et REPERE (Giraudel *et al.*, 2012).

À ces données, s'ajoutent les ensembles MEDIA (Devillers *et al.*, 2004), PORTMEDIA (Lefèvre *et al.*, 2012) et DECODA (Bechet *et al.*, 2012). Ces données sont composées de conversations téléphoniques enregistrées dans un contexte de réservations d'hôtel, de réservations de billet de spectacle et de conversations d'un centre d'appel de la RATP.

La table 1 illustre la quantité de parole de nos ensembles de données.

Nos ensembles de données proviennent de types d'enregistrements différents. Les données issues des émissions de radio et de télévision ont un enregistrement studio en 16 Khz, tandis que les données téléphoniques ont un enregistrement en 8 Khz. Pour nos travaux, nous avons sous-échantillonné les données studio en 8 Khz pour les additionner aux données téléphoniques.

Pour créer notre corpus audio, nous avons respecté les répartitions officielles de tous nos ensembles de données. Il totalise ainsi 26 heures de développement, 55 heures de test et 357 heures d'entraînement.

Corpus	développement	test	entraînement
EPAC	0	9,5	81
ESTER 2	5	6	89
ETAPE	6,5	6,5	19
QUAERO	0	6,5	68,5
REPERE	3	9,5	35,5
MEDIA	1,5	5	17
PORTMEDIA	2	3,5	7
DECODA	8	8,5	40

TABLE 1 – Répartition des corpus audio utilisés. Les quantités d'audio sont exprimées en heures. Ces valeurs sont calculées sur les segments de parole.

4.2 Entités nommées

Les ensembles de données ETAPE et QUAERO possèdent des transcriptions manuelles annotées en entités nommées. Le formalisme QUAERO (Rosset *et al.*, 2011) est utilisé pour cette annotation. Une entité nommée est définie par un type et une valeur, par exemple la valeur « Paris » possède le type « loc.adm.town ». Avec ce formalisme, les types d'entités nommées possèdent une hiérarchie. En reprenant l'exemple de la valeur « Paris », le type principal « loc » est complété par les sous-types « adm » et « town ». À la hiérarchie s'ajoute une annotation en composant. Ils permettent de décrire davantage les entités nommées. Par exemple « firstname », « lastname » pour une entité de type « pers ». Enfin, les annotions d'entités peuvent être imbriquées. Un exemple selon l'annotation Quaero est : « **<pers.ind** le **<func.ind** président **> <pers.ind <firstname** Emmanuel **> <lastname** Macron **> > > ** ».

Pour nos travaux, nous avons simplifié les annotations. Les composants ont été retirés. La hiérarchie des entités nommées est également supprimée. Nous ne conservons que le type principal. Enfin, l'imbrication a été retirée en ne conservant que les annotations les plus proches du niveau mot. Ainsi, nous obtenons huit catégories pour nos annotations : « pers », « func », « org », « loc », « prod », « amount », « time » et « event ». L'exemple précédent devient : « le **<func** président **> <pers** Emmanuel Macron **>** ».

Comme vu dans la section 3, la fonction CTC permet d'apprendre l'alignement entre l'audio et les séquences à produire. Afin d'entraîner notre système, nous ajoutons les frontières des entités nommées dans les séquences de caractères. Ainsi, la séquence servant à apprendre un système de reconnaissance de la parole « le sculpteur césar est mort hier à paris » devient « le sculpteur **<pers** césar **>** est mort **<time** hier **>** à **<loc** paris **>** ». Comme le système utilisé dans cette étude produit des séquences de caractères, les informations de frontières d'entités nommées sont représentés par un caractère. Nous ajoutons dans les étiquettes prédictibles par la fonction CTC huit caractères pour les balises ouvrantes et un caractère pour la balise fermante.

Les ensembles QUAERO et ETAPE représentent en totalité 107 heures d'audio transcrites et annotées manuellement. Pour maximiser notre quantité de données audio annotées en entités nommées, nous avons utilisé le système NeuroNLP2 [2]. Nous entraînons un modèle à l'aide des données manuelles modifiées pour respecter notre annotation simplifiée. Puis nous l'appliquons aux transcriptions manuelles des données audio, d'émission de radio et de TV, n'ayant pas d'annotation en entités nommées. Notre

2. https ://github.com/XuezheMax/NeuroNLP2

augmentation automatique porte sur les corpus EPAC, REPERE, ESTER 2. Nous composons notre corpus audio annoté en entités nommées avec les annotations manuelles et automatiques. Il représente 14,5 heures de développement, 44 heures de test et 293 heures d'apprentissage.

4.3 Concepts sémantiques

L'ensemble de données cible de l'extraction de concepts sémantiques est MEDIA. Il est composé de 1 257 dialogues séparés en trois parties. Une partie développement comprenant 1 300 phrases, une partie test comprenant 3 500 phrases et une partie apprentissage comprenant 17 700 phrases. Cet ensemble de données est annoté selon 76 concepts sémantiques (Bonneau-Maynard *et al.*, 2005). Ces concepts sont par exemple : « chambre-type », « hotel-etat », « sejour-nbNuit », « temps-jour-semaine ». Par observation, un lien peut être fait entre les annotations en concepts sémantiques et les annotations en entités nommées. Il est aisé de rapprocher les concepts « amount » et « nombre-reservation », ou encore de rapprocher « loc » et « localisation-ville ». Nous avons pu observer que les entités nommées « pers », « loc », « amount », « time » et « event » peuvent être reliées aux annotations en concepts sémantiques du corpus MEDIA. Nous considérons que les concepts définis dans ce corpus sont plus précis que les entités nommées.

Nous avons augmenté notre quantité de données annotées en concepts sémantiques en utilisant le corpus PORTMEDIA. Il est composé de 700 dialogues et contient 10 400 phrases. Cet ensemble de données a été produit dans le but d'étudier la portabilité de domaine. Ainsi, il est proche de l'ensemble MEDIA. Il est annoté selon 36 concepts sémantiques, par exemple « nb-billets » ou « type-billet ». Il y a 26 concepts en commun avec l'annotation de MEDIA (« command-tache », « paiement-montant-entier », ...). Les concepts de PORTMEDIA peuvent aussi être rapprochés des concepts d'entités nommées (qui sont toujours considérés plus génériques).

Dans nos travaux, nous exploitons les données de concepts sémantiques de la même manière que les données d'entités nommées. C'est-à-dire en injectant les frontières des concepts directement dans les séquences que le système devra produire.

4.4 Mode étoile

Dans les travaux de Ghannay *et al.* (2018), il a été proposé un mode étoile ayant pour but d'aider le système à se concentrer sur les concepts et leurs valeurs. La mise en œuvre de ce mode consiste à remplacer l'ensemble des caractères en dehors des concepts et de leurs valeurs par une étoile « * ». Ce remplacement s'effectue dans les séquences cibles utilisées pour l'apprentissage du modèle. Ainsi, l'exemple de la section 4.2 devient « * **<pers** césar **>** * **<time** hier **>** * **<loc** paris **>** ».

La fonction de coût CTC donne la même importance à chaque caractère émis composant une séquence. Les éléments de contexte ont donc une importance réduite vis-à-vis des concepts et de leurs valeurs. Ils sont représentés par un seul caractère « * » contre trois caractères ou plus pour les concepts et leurs valeurs (un caractère pour l'ouverture de concept, au minimum un caractère pour la valeur, et un caractère pour la fermeture du concept). Avec cette réduction de l'ensemble du contexte, une erreur faite sur les concepts et valeurs sera nécessairement plus pénalisante pour le système, ce qui lui permettra de concentrer son apprentissage.

5 Chaîne d'apprentissage

Le transfert d'apprentissage rend possible l'exploitation de données issues d'un espace de caractéristiques proches, afin d'apprendre des connaissances a priori facilitant l'entraînement sur une tâche finale (Pan & Yang, 2010). Il est possible d'apprendre un système à partir d'un système existant, plutôt qu'en repartant de zéro. Cette approche a un intérêt dans le cadre d'un manque de données dans un domaine cible. Pour notre étude, l'ensemble d'apprentissage de MEDIA est composé de seulement 17h de parole. Le transfert d'apprentissage permet d'exploiter l'intégralité des données audio à notre disposition pour pré-apprendre un système de reconnaissance de la parole. Ce système pourra ensuite être spécialisé avec les données MEDIA.

L'apprentissage par curriculum repose sur l'introduction ordonnée des différents concepts à apprendre au sein d'un même ensemble. Ce qui permet à un système d'exploiter les concepts plus généraux pour apprendre les concepts plus spécifiques (Bengio *et al.*, 2009). Un système peut converger davantage et plus rapidement en exploitant les exemples d'apprentissage d'un même ensemble du plus simple au plus compliqué. Dans le cas de nos données, les liens entre les entités nommées et les concepts sémantiques nous permettent de considérer l'ordre de complexité suivant : I) Les données audio sont les plus simples. II) Les données annotées en entités nommées sont plus complexes. III) Enfin, les données annotées en concepts sémantiques sont les plus complexes puisque ces concepts sont plus précis que les entités nommées.

Nos travaux ne s'intègrent pas entièrement dans une démarche de curriculum d'apprentissage, puisque nos données ne sont pas issues d'un même ensemble. Ils ont pour objectifs de tirer parti à la fois du transfert d'apprentissage et du curriculum d'apprentissage. Nous réalisons donc un système appris par transfert d'apprentissage et piloté par une stratégie de curriculum. La mise en œuvre d'un tel système nous permet de compenser notre manque de données pour la tâche finale et d'optimiser davantage le modèle obtenu. Nous réalisons une chaîne d'apprentissages successifs dans laquelle nous organisons les corpus de données utilisés du plus générique au plus spécifique.

Entre chaque étape d'apprentissage, nous conservons les poids de l'ensemble du modèle obtenu comme initialisation du réseau pour l'étape suivante. Toutefois, nous réinitialisons totalement la couche haute (softmax) afin de prédire de nouvelles étiquettes. À chaque étape, les étiquettes de sorties du système sont modifiées en fonction des données utilisées.

Dans nos travaux, nous distinguons quatre étapes :

- ASR : Entraînement d'un modèle de reconnaissance de la parole.
- NER : Entraînement d'un modèle de reconnaissance des entités nommées.
- PM+M : Entraînement d'un modèle d'extraction de concepts sémantiques.
- M : Optimisation d'un modèle sur l'ensemble de données MEDIA.

Lors de l'étape « ASR », nous utilisons la totalité des données présentées dans la section 4.1. L'étape « NER » quant à elle est réalisée à l'aide des données augmentées de la section 4.2. Enfin, l'étape « PM+M » correspond à un apprentissage sur l'ensemble augmenté des données décrites dans la section 4.3.

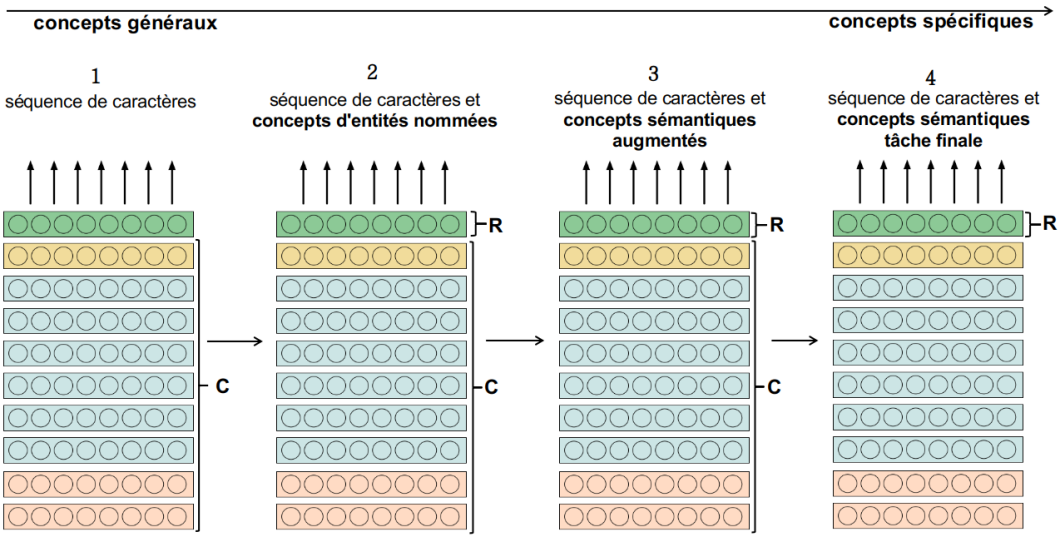

FIGURE 3 – Représentation de la chaîne d'apprentissage mise en place. « C » signifie que les poids sont conservés, « R » signifie qu'ils sont réinitialisés.

6 Expérimentations

Dans cette section, nous présentons les résultats obtenus par nos systèmes sur l'ensemble de données de test de MEDIA. Nos expérimentations sont évaluées selon deux métriques. Le Concept Error Rate (CER), qui consiste en la somme des erreurs de substitution, d'insertion et de suppression divisée par le nombre de références. C'est une métrique proche du Word Error Rate (WER). Cependant, elle permet d'évaluer uniquement la reconnaissance des types de concepts dans les séquences produites, par exemple « chambre-type » / « command-tache ». La seconde métrique utilisée est le Concept Value Error Rate (CVER) qui se calcule comme le CER, mais qui évalue à la fois les types de concepts et leurs valeurs. C'est-à-dire les mots associés aux types de concepts.

L'état de l'art concernant l'extraction de concepts sémantiques au sein de l'ensemble MEDIA se situe à 19,9 % de CER et 25,1 % de CVER (Simonnet *et al.*, 2017). Le système utilisé est une chaîne de composants. Tout d'abord un système de reconnaissance de la parole, puis un système de traitement du langage. Ce dernier exploite les transcriptions automatiques du premier composant. La qualité des transcriptions automatique joue un rôle important dans la qualité finale du système. Les performances du système de transcription de la parole utilisé par l'état de l'art étaient de 23,6 % de WER.

Plusieurs expérimentations contrastives ont été menées afin de vérifier l'impact de chacune des étapes d'apprentissage. Nous avons aussi mené des expérimentations avec des maillons de la chaîne en mode étoile (noté *).

6.1 Mode purement neuronal

chaîne d'apprentissage	CER	CVER
M	39,8	52,1
ASR $\rightarrow$ *M*	23,7	30,3
ASR $\rightarrow$ *NER* $\rightarrow$ *M*	22,4	28,7
ASR $\rightarrow$ *PM+M* $\rightarrow$ *M*	22,2	28,8
ASR $\rightarrow$ *NER* $\rightarrow$ *PM+M* $\rightarrow$ *M*	21,6	27,7
ASR $\rightarrow$ *NER* $\rightarrow$ *PM+M* $\rightarrow$ *M**	**20,1**	27,2
ASR $\rightarrow$ *NER* $\rightarrow$ *PM+M** $\rightarrow$ *M**	**20,1**	**26,9**

TABLE 2 – Résultats expérimentaux exprimés en CER et CVER.

Les résultats expérimentaux montrent l'intérêt des étapes d'apprentissages successives. La table 2 présente les résultats selon la première hypothèse de sortie du système.

L'ensemble de données MEDIA seul n'est clairement pas suffisant pour atteindre de bonnes performances. L'apprentissage d'un système de reconnaissance de la parole comme première étape avant un entraînement sur MEDIA, $ASR \rightarrow M$, permet une amélioration conséquente des performances du système. Les résultats montrent une descente du CER de 39,8 % (M) à 23,7 % ($ASR \rightarrow M$). Cette première étape sert de socle commun à l'ensemble des expérimentations suivantes.

L'utilisation des données d'entités nommées dans la chaîne $ASR \rightarrow NER \rightarrow M$ montre une utilité en raison d'une amélioration significative de 1,3 point de CER, par rapport à la chaîne $ASR \rightarrow M$. Nous atteignons un CER de 22,4 %, soit un gain relatif de 5,5 %. L'exploitation de l'augmentation des données MEDIA avec l'ensemble PORTMEDIA nous permet aussi un gain en descendant de 23,7 % à 22,2 % de CER (comparaison des chaînes $ASR \rightarrow M$ et $ASR \rightarrow PM + M \rightarrow M$). Enfin, nous avons appris un modèle exploitant pleinement le transfert d'apprentissage piloté par la stratégie de curriculum proposé dans ce papier. Ce modèle, $ASR \rightarrow NER \rightarrow PM + M \rightarrow M$, atteint 21,6 % de CER. Il montre à nouveau l'utilité des données d'entités nommées, par une amélioration de 0,6 point. Soit un gain relatif de 2,7 %.

L'utilisation du mode étoile, décrit dans la section 4.4, aide le système à ce concentrer sur les concepts et leurs valeurs. Ce qui nous permet d'améliorer une dernière fois le CER de 1,5 %. Nous atteignons une valeur de 20,1 % (CER).

Toutes les observations que nous avons réalisées sur les taux de CER peuvent être effectuées sur les taux de CVER.

6.2 Mode à double étape avec modèle de langage n-gramme

Le système deepspeech 2 permet de recalculer ses sorties avec un modèle de langage. Le modèle de langage utilisé est au niveau mot. Il est appris à l'aide de l'ensemble d'apprentissage de MEDIA, ainsi que d'un ensemble conséquent de données issues d'articles de journaux. Les informations de concepts sémantiques, encodées sur un caractère, sont conservées pour l'entraînement du modèle de langage. Un second est appris avec le mode étoile. Lors de nos expérimentations, nous avons calculé des modèles variant de 3-gramme à 6-gramme. Nous utilisons les modèles 5-gramme, puisque nous n'observons plus d'amélioration significative au-delà. Les résultats de nos expérimentations avec les modèles de langage sont reportés dans la table 3.

chaîne d'apprentissage	CER	CVER
M	32,8	37,9
ASR $\rightarrow$ *M*	20,1	24,0
ASR $\rightarrow$ *NER* $\rightarrow$ *M*	18,8	22,8
ASR $\rightarrow$ *PM+M* $\rightarrow$ *M*	19,0	22,9
ASR $\rightarrow$ *NER* $\rightarrow$ *PM+M* $\rightarrow$ *M*	18,1	22,1
ASR $\rightarrow$ *NER* $\rightarrow$ *PM+M* $\rightarrow$ *M**	16,6	21,3
ASR $\rightarrow$ *NER* $\rightarrow$ *PM+M** $\rightarrow$ *M**	**16,4**	**20,9**

TABLE 3 – Résultats expérimentaux exprimés en CER et CVER. Les résultats sont recalculés à l'aide d'un modèle de langage 5-gramme.

L'utilisation des modèles de langage nous permet une amélioration significative des résultats sur l'ensemble de nos expériences. L'intérêt des entités nommées dans notre chaîne d'apprentissage est conservé. En comparant les systèmes $ASR \rightarrow M$ et $ASR \rightarrow NER \rightarrow M$, nous observons un gain relatif de 6,5 % de CER et en comparant les systèmes $ASR \rightarrow PM+M \rightarrow M$ et $ASR \rightarrow NER \rightarrow PM+M \rightarrow M$, il est de 4,7 %.

Par l'utilisation du mode étoile, nous obtenons désormais nos meilleurs taux de CER et de CVER, avec respectivement 16,4 % et 20,9 %. Pour obtenir ces résultats, nous avons utilisé le mode étoile pendant les étapes d'apprentissage manipulant les concepts sémantiques. L'application du mode étoile aux données d'entités nommées a pour effet de dégrader nos résultats. Nous observons que le mode étoile doit être appliqué au niveau des étapes d'extraction de concepts sémantiques. En comparant nos meilleurs résultats à l'état de l'art (CER : 19,9 % et CVER : 25,1 %), nous obtenons de très bons résultats. Nous avons un gain relatif de 17,6 % pour le CER, et de 16,7 % pour le CVER.

Le système que nous avons mis en œuvre obtient un WER de 10,1 %. Il est calculé sur les sorties de la chaîne d'apprentissage $ASR \rightarrow NER \rightarrow PM+M \rightarrow M$. Ces sorties sont recalculées avec le modèle de langage 5-gramme, puis toutes les informations de concepts sémantiques sont filtrées pour évaluer le WER sur les mots uniquement.

Nous obtenons dans notre approche un WER bien meilleur que l'état de l'art, 10,1 % contre 23,6 % (soit 57,2 % de gain relatif). La comparaison de nos résultats n'est pas entièrement juste et il conviendrait de mettre en œuvre le système à chaîne de composants de l'état de l'art avec des transcriptions automatiques d'une qualité similaire à notre système.

6.3 Validation de la stratégie de curriculum

Pour compléter nos expérimentations, nous avons tenté l'apprentissage de la chaîne $ASR \rightarrow PM+M \rightarrow NER \rightarrow M$, qui rompt le processus itératif de spécialisation des différentes étapes d'apprentissage en inversant les étapes NER et $PM+M$. Nous observons lors de l'apprentissage de l'étape NER que le modèle ne converge pas, et ce malgré nos différentes tentatives de modification des paramètres d'apprentissage, notamment le taux d'apprentissage. Cette absence de convergence confirme l'apport de l'ordonnancement des étapes d'apprentissage que nous avons présenté dans notre stratégie de curriculum. Ainsi, en ayant inversé les étapes NER et $PM+M$, nous ne pouvons pas tirer parti des connaissances portées par les données étiquetées en entités nommées.

7 Conclusion

Ce travail présente l'intérêt de l'utilisation des données d'entités nommées pour la reconnaissance de concepts sémantiques à travers la mise en place d'une chaîne d'apprentissages successifs pilotée par une stratégie de curriculum. Nous avons réalisé des augmentations de données via un ensemble proche (PORTMEDIA) pour les concepts sémantiques et via une augmentation artificielle (NeuroNLP2) pour les entités nommées. Nos données viennent alimenter le système DeepSpeech 2. Pour ce faire, nous avons intégré les frontières des entités nommées et des concepts sémantiques directement dans les chaînes de caractères à produire. Les données sont présentées au système de manière à apprendre une tâche de reconnaissance de la parole, puis une tâche de reconnaissance des entités nommées dans la parole et enfin une tâche d'extraction de concepts sémantiques.

Les résultats expérimentaux montrent une amélioration des performances de cette approche, qui permet un gain relatif allant de 2,7 % à 6,5 % de Concept Error Rate. Notre meilleur système est obtenu en exploitant pleinement la chaîne d'apprentissage pilotée par la stratégie de curriculum proposée dans ce papier. Ce système utilise également un mode étoile, lui permettant de se concentrer sur les concepts et leurs valeurs, ainsi qu'un modèle de langage 5-gramme pour recalculer ses sorties. Il surpasse l'état de l'art sur la tâche d'extraction de concepts sémantiques dans MEDIA, avec un gain relatif de 17,6 % de Concept Error Rate.

Il convient cependant de nuancer nos résultats, puisque notre système obtient de bien meilleures performances en termes de reconnaissance de la parole que le composant parole du système état de l'art. Nos expérimentations ont montré un gain relatif de 57,2% de WER. Un complément nécessaire à ce travail consiste en l'utilisation de transcriptions automatiques ayant des performances similaires à notre approche avec le système état de l'art.

En conclusion, ces travaux présentent des premiers résultats intéressants à propos de l'intégration des entités nommées dans une chaîne d'apprentissage, basée sur les principes de transfert et de curriculum d'apprentissage, pour l'extraction de concepts sémantiques. Ils constituent un socle intéressant pour de futurs travaux visant à approfondir les possibilités de l'exploitation des entités nommées comme données génériques, notamment pour la portabilité de domaine. Ce socle peut également être amélioré par l'ajout d'informations supplémentaires pouvant aider l'extraction de concepts sémantiques. En effet, nos expérimentations se basent uniquement sur les mots et les concepts pour effectuer la tâche finale. Nous pouvons envisager de profiter des années de travaux effectuer dans le domaine du traitement du langage pour enrichir nos données d'apprentissage, comme par exemple l'utilisation d'étiqueteurs morpho-syntaxiques ou sémantique.

Remerciements

Ce travail a été soutenu par le RFI Atlanstic2020 à travers le projet RAPACE (Réseaux de neurones profonds pour le traitement de la langue orale et écrite). Il a également été soutenu par l'agence ANR au travers du projet CHIST-ERA ON-TRAC, sous le numéro de contrat : ANR-18-CE23-0021-01. Les auteurs souhaitent remercier Sean Naren pour la mise à disposition de son implémentation du système DeepSpeech 2, ainsi que Xuezhe Ma pour son implémentation du système NeuroNlp2.

Références

AMODEI D., ANANTHANARAYANAN S., ANUBHAI R., BAI J., BATTENBERG E., CASE C., CASPER J., CATANZARO B., CHENG Q., CHEN G. *et al.* (2016). Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of the thirty-third International Conference on Machine Learning (ICML'16)*, p. 173–182, New York, United States.

BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BEZE M., DE MORI R. & ARBILLOT E. (2012). Decoda : a call-centre human-human spoken conversation corpus. In *Proceedings of the eighth Language Resources and Evaluation Conference (LREC'12)*, p. 1343–1347, Istanbul, Turkey.

BENGIO Y., LOURADOUR J., COLLOBERT R. & WESTON J. (2009). Curriculum learning. In *Proceedings of the twenty-sixth International Conference on Machine Learning (ICML'09)*, p. 41–48, Montreal, Canada.

BÉRARD A., BESACIER L., KOCABIYIKOGLU A. C. & PIETQUIN O. (2018). End-to-end automatic speech translation of audiobooks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'18)*, p. 6224–6228, Calgary, Canada.

BÉRARD A., PIETQUIN O., BESACIER L. & SERVAN C. (2016). Listen and translate : A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on End-to-end Learning for Speech and Audio Processing*, Barcelona, Spain.

BONNEAU-MAYNARD H., ROSSET S., AYACHE C., KUHN A. & MOSTEFA D. (2005). Semantic annotation of the french media dialog corpus. In *Proceedings of the ninth European Conference on Speech Communication and Technology (EUROSPEECH'05)*, p. 3456–3459, Lisbon, Portugal.

DEVILLERS L., MAYNARD H., ROSSET S., PAROUBEK P., MCTAIT K., MOSTEFA D., CHOUKRI K., CHARNAY L., BOUSQUET C., VIGOUROUX N. *et al.* (2004). The french media/evalda project : the evaluation of the understanding capability of spoken language dialogue systems. In *Proceedings of the fourth Language Resources and Evaluation Conference (LREC'04)*, p. 2131–2134, Lisbon, Portugal.

ESTEVE Y., BAZILLON T., ANTOINE J.-Y., BÉCHET F. & FARINAS J. (2010). The epac corpus : Manual and automatic annotations of conversational speech in french broadcast news. In *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC'10)*, p. 1686–1689, Malta.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Proceedings of the tenth Conference of the International Speech Communication Association (INTERSPEECH'09)*, p. 2543–2546, Brighton, United Kingdom.

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 *Proceedings of the IEEE Spoken Language Technology Workshop (SLT'19)*, p. 692–699, Athens, Greece.

GIRAUDEL A., CARRÉ M., MAPELLI V., KAHN J., GALIBERT O. & QUINTARD L. (2012). The repere corpus : a multimodal corpus for person recognition. In *Proceedings of the eighth Language Resources and Evaluation Conference (LREC'12)*, p. 1102–1107, Istanbul, Turkey.

GRAVES A., FERNÁNDEZ S., GOMEZ F. & SCHMIDHUBER J. (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the twenty-third International Conference on Machine Learning (ICML'06)*, p. 369–376, Pittsburgh, United States.

GRAVIER G., ADDA G., PAULSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. In *Proceedings of the eighth Language Resources and Evaluation Conference (LREC'12)*, p. 114–118, Istanbul, Turkey.

GROUIN C., ROSSET S., ZWEIGENBAUM P., FORT K., GALIBERT O. & QUINTARD L. (2011). Proposal for an extension of traditional named entities : From guidelines to evaluation, an overview. In *Proceedings of the fifth Linguistic Annotation Workshop*, p. 92–100, Portland, United States.

HANNUN A. (2017). Sequence modeling with ctc. *Distill*. https ://distill.pub/2017/ctc.

JAN N., CATTONI R., SEBASTIAN S., CETTOLO M., TURCHI M. & FEDERICO M. (2018). The iwslt 2018 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, p. 2–6, Bruges, Belgium.

LEFÈVRE F., MOSTEFA D., BESACIER L., ESTÈVE Y., QUIGNARD M., CAMELIN N., FAVRE B., JABAIAN B. & ROJAS-BARAHONA L. (2012). Robustesse et portabilités multilingue et multi-domaines des systèmes de compréhension de la parole : les corpus du projet portmedia. In *Proceedings of the Joint Conference JEP-TALN-RECITAL*, p. 779–786, Grenoble, France.

PAN S. J. & YANG Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, p. 1345–1359.

PLATANIOS E. A., STRETCU O., NEUBIG G., POCZOS B. & MITCHELL T. M. (2019). Competence-based curriculum learning for neural machine translation. *arXiv preprint arXiv :1903.09848*.

ROSSET S., GROUIN C. & ZWEIGENBAUM P. (2011). *Entités nommées structurées : guide d'annotation Quaero*. LIMSI-Centre national de la recherche scientifique.

SERDYUK D., WANG Y., FUEGEN C., KUMAR A., LIU B. & BENGIO Y. (2018). Towards end-to-end spoken language understanding. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'18)*, p. 5754–5758, Calgary, Canada.

SIMONNET E., GHANNAY S., CAMELIN N., ESTÈVE Y. & DE MORI R. (2017). ASR error management for improving spoken language understanding. In *Proceedings of the eight-teenth Conference of the International Speech Communication Association (INTERSPEECH'17)*, p. 3329–3333, Stockholm, Sweden.

WEINSHALL D., COHEN G. & AMIR D. (2018). Curriculum learning by transfer learning : Theory and experiments with deep networks. *arXiv preprint arXiv :1802.03796*.

WEISS R. J., CHOROWSKI J., JAITLY N., WU Y. & CHEN Z. (2017). Sequence-to-sequence models can directly translate foreign speech. In *Proceedings of the eight-teenth Conference of the International Speech Communication Association (INTERSPEECH'17)*, p. 2625–2629, Stockholm, Sweden.