Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Nouvelle tâche sémantique pour le corpus de compréhension de parole en français MEDIA

Nadège Alavoine[1] Gaëlle Laperrière[2] Christophe Servan[1,2]
Sahar Ghannay[1] Sophie Rosset[1]
[1]Université Paris-Saclay, CNRS, LISN, Campus Universitaire bât,507 - Rue du Belvédère - 91405 Orsay
[2]Avignon Université, LIA, 339 chemin des Meinajariés - BP 1228 - 84911 Avignon Cedex 9
[3]QWANT, 10 bvd Haussmann - Paris 75009
{firstname.lastname}@lisn.upsaclay.fr

RÉSUMÉ

La détection d'intention et de concepts sont des tâches essentielles de la compréhension de la parole (SLU). Or il n'existe que peu de données annotées en français permettant d'effectuer ces deux tâches conjointement. Cependant, il existe des ensembles de données annotées en concept, dont le corpus MEDIA. Ce corpus est considéré comme l'un des plus difficiles. Néanmoins, il ne comporte que des annotations en concepts et pas en intentions. Dans cet article, nous proposons une version étendue de MEDIA annotée en intentions pour étendre son utilisation. Cet article présente une méthode semi-automatique pour obtenir cette version étendue. De plus, nous présentons les premiers résultats des expériences menées sur cet ensemble de données en utilisant des modèles joints pour la classification des intentions et la détection de concepts.

ABSTRACT

New Semantic Task for the French Spoken Language Understanding MEDIA Benchmark

Intention and concepts detection are essential tasks in speech understanding (SLU). There are a few annotated data sets in French, both in concepts and intention. However, there are some French datasets annotated in concept, including MEDIA. This French dataset, distributed since 2005 by ELRA, is one of the top SLU task. Unfortunately, it is only annotated in concepts and not in intent. In this article, we propose an improved version of MEDIA annotated with intentions to extend its use. This article presents the semi-automatic methodology used to obtain this improved version. In addition, we present the first results of experiments on this improved dataset using joint models for intention classification and concepts detection.

MOTS-CLÉS : Données d'évaluation, compréhension de la parole, détection jointe de l'intention et de concepts, tri-apprentissage.

KEYWORDS: Benchmark Dataset, Spoken Language Understanding, Joint Intent Detection And Slot-filling, Tri-training.

1 Introduction

Le module de compréhension de la parole (en anglais *Spoken Language Understanding* – SLU) est un élément crucial d'un système de dialogue oral. Les tâches de SLU regroupent trois sous-tâches : la classification de domaine, la détection d'intentions, et l'annotation de séquences en concepts sémantique (ou détection de concepts) (Tur & Mori, 2011). Dans cette étude, nous nous sommes

Actes des 35ème Journées d'Études sur la Parole, pages 470–480.

intéressés à la détection d'intentions et à la tâche de remplissage de formulaire. Cette dernière tâche peut également être considérée comme une tâche de détection de concepts (Bonneau-Maynard *et al.*, 2006).

La plupart des systèmes de dialogue traitent ces tâches séparément en développant des modules indépendants insérés dans un pipeline (Hakkani-Tür *et al.*, 2016). Ces approches pipeline souffrent généralement de la propagation d'erreurs en raison de leurs modèles indépendants. Ainsi, des modèles joints de classification des intentions et de détection de concepts ont été proposés pour résoudre ce problème et pour améliorer mutuellement ces deux tâches (Weld *et al.*, 2022). Pour ces modèles joints, plusieurs approches ont été explorées, telles que, des modèles fondés sur des champs conditionnels aléatoires (CRF) (Jeong & Lee, 2008), des réseaux de neurones convolutionels (Xu & Sarikaya, 2013), récurrents (Guo *et al.*, 2014; Hakkani-Tür *et al.*, 2016; Liu & Lane, 2016), avec des portes de remplissages (*slot-gated models*) (Goo *et al.*, 2018), avec mécanisme d'attention (Chen *et al.*, 2016; Hakkani-Tür *et al.*, 2016; Liu & Lane, 2016), et des modèles Transformers pré-entraînés (Chen *et al.*, 2019; Castellucci *et al.*, 2019; Wang *et al.*, 2020; Qin *et al.*, 2021; Han *et al.*, 2021) ou des modèles convolutionels graphiques (Tang *et al.*, 2020). Pour l'anglais, les modèles joints sont traditionnellement évalués sur des tâches annotées avec des intentions et de concepts : ATIS (Hemphill *et al.*, 1990) et SNIPS (Coucke *et al.*, 2018). Cependant, en français, moins de ressources sont disponibles. Il existe des données ATIS étendues au français dans le corpus MultiATIS++ (Xu *et al.*, 2020) par traduction. La version étendue MultiATIS++ ne dispose malheureusement pas des supports audios correspondants aux nouvelles langues. Les données MEDIA comportent les supports audios. Malheureusement, ces données ne sont annotées qu'en concepts et pas en intention, même si en l'état ce corpus est considéré comme l'un des plus difficiles (Béchet & Raymond, 2019).

Cet article présente une version mise-à-jour des données MEDIA avec des annotations d'intention à l'aide d'une approche semi-automatique. De plus, nous présentons les premiers résultats des expériences de compréhension sur cet ensemble de données amélioré à l'aide de modèles conjoints pour la classification des intentions et la détection de concepts.

2 Annotation du corpus MEDIA en intentions

Le corpus MEDIA est composé d'appels téléphoniques enregistrés pour la réservation d'hôtel. Il est dédié à l'extraction sémantique de l'information de la parole dans le contexte des dialogues homme-machine recueillis en utilisant la méthode Wizard-of-Oz (Bonneau-Maynard *et al.*, 2005). L'ensemble de données représente 1258 dialogues enregistrés de 250 différents locuteurs et environ 70 heures de conversations. Seuls les tours de parole des utilisateurs sont annotés avec des transcriptions manuelles et des annotations sémantiques complexes (concepts), et utilisés dans ce travail. Le corpus MEDIA est disponible en version *full* ou *relax*. Dans le second, les attributs sont simplifiés en excluant les spécifieurs. Récemment, Laperrière *et al.* (2022) ont proposé une version mise-à-jour de l'ensemble de données MEDIA. Le travail présenté ici est fondé sur cette version de MEDIA que nous notons MEDIA 2022 par la suite.

À notre connaissance, le corpus MEDIA n'a jamais été annoté avec des intentions. Contrairement à d'autres ensembles de données de référence comme ATIS (Hemphill *et al.*, 1990) ou SNIPS (Coucke *et al.*, 2018), seule la détection de concepts a été prise en compte. Nous proposons ici une version étendue du corpus MEDIA 2022 annotée en intentions. Pour cela, nous avons défini une liste de 11 intentions après avoir examiné attentivement le contenu de l'ensemble de données. Un tour de parole peut recevoir plusieurs intentions. Dans un tel cas, les labels sont séparées par le signe #. Les détails de cette liste, des exemples et des contre-exemples sont disponibles dans un guide d'annotation.

2.1 Méthodologie

L'annotation d'un ensemble de données peut être une tâche chronophage. Pour réduire le coût et le temps d'annotation, nous avons utilisé une méthode de tri-apprentissage (Zhou & Li, 2005) pour augmenter la taille des données d'apprentissage. Le tri-apprentissage est une méthode inductive épisodique semi-supervisée (van Engelen & Hoos, 2020) visant à améliorer les performances de n'importe quel type de système en ajoutant des données non-étiquetées. Il utilise un trio de classifieurs formés sur différents ensembles de données d'apprentissage. A chaque épisode de l'algorithme, ces classifieurs attribuent une *pseudo-étiquette* (Chen *et al.*, 2019) à des données non-étiquetées. Lorsque deux classifieurs du trio s'accordent sur une *pseudo-étiquette*, les données correspondantes *pseudo-étiquetées* sont ajoutées à l'ensemble d'apprentissage du troisième modèle. Lors de l'épisode suivant, les classifieurs pourront poursuivre leur apprentissage sur les ensembles de données mis à jour. L'algorithme de tri-apprentissage s'arrête quand aucun changement ne peut être observé dans l'apprentissage de tous les classifieurs du trio.

Récemment, Boulanger *et al.* (2022) ont montré que le tri-apprentissage pouvait être utilisé dans un cadre de faible ressource sur une tâche de reconnaissance d'entité nommée (REN). Nous avons décidé d'utiliser un système similaire pour entrainer et évaluer des trios de classifieurs. Le meilleur trio sera retenu pour annoter l'ensemble du corpus MEDIA en intention.

2.1.1 Ensembles de données pour le tri-apprentissage

Une partie des données annotées manuellement est nécessaire pour l'apprentissage et l'évaluation des trios de classifieurs. À cette fin, nous avons utilisé une version transcrite de l'ensemble de données MEDIA résultant du système de segmentation le plus couramment utilisé, avec des mots tronqués conservés comme mots entiers ("mer(ci)" est écrit "merci"). Cette version sera notée "MEDIA original". Un sous-ensemble d'énoncés choisis au hasard à partir de l'ensemble du corpus d'apprentissage original et d'autres choisis explicitement pour leur contenu ont été annotés manuellement en suivant notre guide d'annotation (disponible en ligne dans le dépôt). Cette annotation a été réalisée hors contexte : chaque déclaration a été traitée sans tenir compte des précédents énoncés dans le dialogue. 1551 énoncés ont été annotés manuellement pour le tri-apprentissage, 1240 constituant le corpus d'apprentissage, 124 pour le corpus de développement et 187 pour le corpus d'évaluation.

	Tri-apprentissage				**MEDIA**			
Intentions	**Appr.**	**Val.**	**Éval.**	**Total**	**Appr.**	**Val.**	**Éval.**	**Total**
annulation	15	1	1	17	32	1	15	48
incompréhension	6	1	4	11	273	30	94	397
marqueur_de_discours	38	6	5	49	282	40	113	435
modification	7	1	1	9	115	10	31	156
merci	47	5	6	58	713	100	200	1013
information	114	11	19	144	1611	159	401	2171
réponse_oui	392	42	52	486	4325	419	1190	5934
réponse_indécis	9	1	3	13	37	5	9	51
réponse_non	362	35	57	454	1315	88	344	1747
réservation	352	30	48	430	5437	522	1410	7369
salutation	43	8	6	57	717	101	206	1024

TABLE 1 – Distribution des étiquettes dans un sous-échantillonage du corpus MEDIA utilisé pour le tri-apprentissage (partie de gauche) et sur l'ensemble de données MEDIA (partie de droite). Ces sous-ensembles sont découpés en apprentissage (Appr.), validation (Val.) et évaluation (Éval.).

2.1.2 Protocole expérimental

Notre cas d'utilisation diffère du travail de Boulanger *et al.* (2022), car nous avons beaucoup de données non annotées. Nous avons adapté leur code en désactivant la génération de données synthétiques

et en modifiant le classificateur pour réaliser une pseudo-annotation *multi-label*. Il utilise l'état caché final du jeton spécial [CLS] combiné avec une couche sigmoïde et un seuil de $0,5$ pour déterminer l'intention à associer avec la phrase.

Nous avons utilisé deux modèles français de Transformers (Vaswani *et al.*, 2017) : CamemBERT (Martin *et al.*, 2020), un modèle dérivé de RoBERTA (Zhuang *et al.*, 2021), et FrALBERT (Cattan *et al.*, 2021), un modèle compact dérivé d'ALBERT (Lan *et al.*, 2020). Nous avons évalué deux versions comparables, formées sur 4 gigabytes (GB) de texte à partir du site Web de Wikipedia : CamemBERT-base-Wikipedia-4GB [1]et FrALBERT-base [2]. Cattan *et al.* (2022) ont montré que les classifieurs basés sur ces modèles avaient de bonnes performances en SLU sur l'ensemble de données d'évaluation MEDIA pour la tâche de détection de concepts (*slot-filling*).

Avant l'algorithme de tri-apprentissage, un échantillonnage aléatoire de 1000 énoncés parmi les 1240 qui constituent notre corpus d'apprentissage de tri-apprentissage est effectué pour chaque modèle de trio. Le réglage fin (*fine-tuning*) de nos modèles sur cette portion de données diminue les chances que les trois classifieurs produisent les mêmes résultats. L'algorithme est entraîné sur un maximum de 30 époques, bien qu'il s'arrête une fois qu'aucune variation du score d'évaluation n'est observée sur le corpus de validation. Les hyperparamètres sont fixés avec un taux d'apprentissage de 1e-5, une taille de lot d'apprentissage de 16 et une valeur de dropout de $0,1$. Le nombre maximal d'époques par épisode est de 1000, avec une méthode d'arrêt précoce de 20 époques.

Les performances des classifieurs sont évaluées au cours de l'apprentissage avec le rapport de correspondance exact (*Exact Match Ratio*, abrégé EMR) (Sorower, 2010) d'intention sur l'ensemble de validation. Une fois l'algorithme de tri-apprentissage arrêté, EMR, précision, rappel et F-mesure (ou score F1) sont évalués sur l'ensemble d'évaluation présenté dans le Tableau 1 (partie de gauche). Ces performances sont calculées sur les votes majoritaires des prédictions de trios de modèles.

2.1.3 Évaluation

Les résultats de nos expériences utilisant les données présentées dans la partie gauche du Tableau 1 sont présentés dans le Tableau 2. La plupart des expériences se sont arrêtées après 3 ou 4 époques de tri-apprentissage. Les trios utilisant le modèle CamemBERT obtiennent de meilleurs résultats que les trios utilisant FrALBERT en les dépassant de $7,17$ points sur le EMR et de $5,09$ points sur la F-mesure. Ils ont également moins de variabilité dans leurs résultats, avec un écart type oscillant entre $0,33$ et $0,70$ sur les différentes mesures par rapport à $0,81$ à $1,62$ pour FrALBERT.

Transformer	**EMR**	**Précision**	**Rappel**	**F1**
CamemBERT	92,09 $\pm 0,45$	95,29 $\pm 0,70$	93,48 $\pm 0,36$	93,73 $\pm 0,33$
FrALBERT	84,92 $\pm 1,62$	90,86 $\pm 0,81$	87,97 $\pm 1,44$	88,64 $\pm 1,37$

TABLE 2 – Résultats de classification en intentions avec l'algorithme de tri-apprentissage en utilisant une partie de l'ensemble de données MEDIA.

Suite à ces résultats, nous avons examiné de plus près les performances de notre meilleur trio de modèles de CamemBERT. Ce trio sera retenu pour annoter automatiquement l'ensemble de données MEDIA. Le trio obtient un EMR de $92,51$ points et une F-mesure de $93,85$ points par échantillon. En ce qui concerne la performance de la macro F-mesure, elle est à $58,99$ points. Cette macro F-mesure semble fortement influencée par une proportion importante de faux négatifs dans certaines

1. https ://huggingface.co/CamemBERT/CamemBERT-base-Wikipedia-4GB
2. https ://huggingface.co/qwant/FrALBERT-base

étiquettes, avec un macro rappel de $60,98$ points. En revanche, les faux positifs sont rares, avec une macro précision de $93,77$ points. Les faux négatifs concernent principalement les étiquettes avec peu d'exemples dans notre jeu d'évaluation présenté dans le Tableau 1 car ils n'affectent pas la moyenne de l'échantillon de rappel et de F-mesure autant.

2.1.4 Discussion, annotations, et corrections

Ce travail représente une première approche vers l'utilisation d'un algorithme de tri-apprentissage avec des classifieurs basés sur les Transformers pour annoter un ensemble de données. Puisque notre but était d'accélérer l'annotation, nous avons gardé la *pseudo-étiquette* pour lequel notre meilleure trio a obtenu un consensus. Pour chaque combinaison de *pseudo-étiquettes*, les énoncés correspondants ont été présentés à l'annotateur, qui a dû invalider les tentatives erronées. Les énoncés avec *pseudo-étiquettes* erronées ou sans aucune *pseudo-étiquette* attribuée, faute de consensus du trio de modèles, ont été annotés manuellement. Il y a eu 3122 intentions totalement ou partiellement erronées ($19,51$ % des données de 16005 *pseudo-étiquetées* et 137 de non-*pseudo-étiquetées*). L'ensemble des intentions étiquetées sur le corpus MEDIA sont présentées dans la partie de droite du Tableau 1.

2.2 Annotation de la version MEDIA 2022

La version MEDIA 2022 a également été annotée. Pour les ensembles d'apprentissage, validation et d'évaluation, la méthodologie utilisée diffère de celle décrite dans la section 2.1 puisque nous avions déjà les intentions associées à chaque énoncé. Une correspondance sur le contenu textuel des énoncés a été réalisée pour récupérer les annotations dans la mesure du possible.

3 Expériences sur les transcriptions manuelles

En utilisant l'ensemble enrichi des données MEDIA, nous présentons une première évaluation sur les transcriptions manuelles, en effectuant l'apprentissage joint des tâches de détection d'intentions et de concepts.

3.1 Architecture neuronale

Le modèle joint de détection d'intention et de de concepts utilisé dans nos expériences est le modèle JointBERT (Chen *et al.*, 2019). Pour la tâche de détection des intentions, ce modèle combine l'état caché final du jeton [CLS] à une couche *softmax*. Pour la détection de concepts, il détermine quel concept peut être associé à chaque mot en fournissant l'état final de chaque première sous-position d'un mot à une couche *softmax*. Le modèle est affiné en optimisant la somme des pertes d'entropie croisée pour les deux tâches.

La probabilité P_i pour qu'une phrase soit associée à une intention i passant $h_{[CLS]}$ (le dernier état caché du transformeur pour le jeton[CLS]) à une couche de poids W^i et de biais b^i est définie comme suit : $P_i = sigmoide(W^i h_{[CLS]} + b^i) > 0,5$. Une perte d'entropie croisée binaire remplace la perte d'entropie croisée précédemment utilisée pour la classification des intentions. Le modèle est optimisé sur la somme des pertes d'entropie croisée binaires et non binaires pour la classification d'intentions et la détection de concepts respectivement.

3.2 Protocole expérimental

Pour la tâche de détection de concepts, nous avons utilisé un format BIO. Les performances sont évaluées en termes de micro F-mesure, couramment utilisée pour les modèles joints (Weld *et al.*, 2022), et de Concept Error Rate (CER) qui est la mesure officielle utilisée dans la campagne MEDIA (Bonneau-Maynard *et al.*, 2006). Pour une comparaison avec les expériences sur les sorties de reconnaissance automatique de la parole (RAP), nous utilisons la micro F-mesure calculée sur des vecteurs *multi-hot* (abrégée F1mh) de concepts présents dans les annotations attendues et obtenues.

Pour la classification des intentions, lorsqu'il y a plusieurs intentions, nous les concaténons en utilisant le caractère (#). Dans la plupart des modèles conjoints, la performance de cette tâche est évaluée à l'aide de l'exactitude (*accuracy* en anglais, abrégée Exa.) (Weld *et al.*, 2022). Comme nous utilisons un système de classification *multi-label*, l'exactitude proposée par Godbole & Sarawagi (2004) et EMR ont été évaluées. L'exactitude du cadre sémantique de la phrase (*sentence-level semantic frame accuracy*, abrégée SFA) - correspondant au nombre d'énoncés avec une intention et des concepts correctement trouvés divisés par le nombre de phrases - couramment utilisée pour les modèles joints (Weld *et al.*, 2022), est également évaluée.

Nous avons choisi le modèle de base CamemBERT entrainé sur 135 GB de texte de CCNET (CamemBERT-base-CCNET) (Martin *et al.*, 2020) ainsi que les modèles de base CamemBERT-base-Wikipedia-4GB et FrALBERT utilisés précédemment. Ces modèles ont montré des résultats à l'état-de-l'art , ou proches de ceux-ci, pour la détection de concepts sur les transcriptions manuelles MEDIA (Ghannay *et al.*, 2020; Cattan *et al.*, 2022). Nous avons également choisi un modèle français BERT, FlauBERT, optimisé pour quelques époques sur des données transcrites par RAP (FlauBERT-oral-ft) (Hervé *et al.*, 2022) qui a également obtenu des performances proches de l'état-de-l'art sur les sorties MEDIA ASR (Pelloin *et al.*, 2022).

De la même manière que Cattan *et al.* (2022), nous avons utilisé un algorithme génétique, le *population based training* (abrégé PBT) (Jaderberg *et al.*, 2017), pour déterminer les meilleurs hyperparamètres.

3.3 Résultats sur les transcriptions manuelles

Les performances sur la version originale et relax de MEDIA, sur la version relax de MEDIA 2022, et la version full de MEDIA 2022 sont affichées dans la partie gauche du Tableau 3.

En ce qui concerne les scores de la tâche de détection de concepts, nous pouvons logiquement observer que les modèles obtiennent de meilleurs résultats sur la version relax que la version full. Par exemple, sur MEDIA 2022, il y a une différence de $2,42$ points de F-mesure pour les meilleurs résultats obtenus entre la version relax et la version full, en faveur de la version relax. Plus surprenamment, tous les modèles fonctionnent mieux sur les deux tâches avec la version relax d'origine que sur la version MEDIA 2022 relax. Cela pourrait s'expliquer par la conservation de mots tronqués dans MEDIA 2022, qui compliquerait ces tâches.

4 Expériences sur la reconnaissance automatique de la parole

Nous évaluons les performances des deux approches à l'aide des mesures utilisées dans la section 3.2, à l'exception de la F-mesure de la détection de concepts.

	Transcriptions manuelles						Cascade avec RAP				
	Intentions		Concepts				Intentions		Concepts		
Modèle	**Exa.**	**EMR**	**F1**	**F1mh**	**CER**	**SFA**	**Exa.**	**EMR**	**F1mh**	**CER**	**SFA**
MEDIA original, relax											
CamemBERT-base-CCNET	93,87	91,79	88,52	**95,97**	8,68	**76,26**	92,07	89,82	**93,82**	13,93	**65,69**
CamemBERT-base-Wikipedia-4GB	**93,98**	**91,84**	87,93	95,41	9,34	75,58	**92,43**	**90,08**	93,29	15,03	65,49
FlauBERT-oral-ft	93,66	91,19	87,93	95,63	8,95	76,04	92,28	89,77	93,12	14,19	65,55
Base FrALBERT	92,27	89,88	84,24	93,66	13,14	72,12	90,81	88,18	91,14	19,97	62,91
MEDIA 2022**, relax**											
CamemBERT-base-CCNET	91,87	**89,78**	86,95	94,66	10,33	72,68	90,16	**88,00**	92,74	12,78	64,43
CamemBERT-base-Wikipedia-4GB	91,25	88,66	86,88	94,88	10,24	72,60	89,39	86,75	92,90	13,19	64,00
FlauBERT-oral-ft	**92,10**	89,73	**87,75**	**95,41**	**9,18**	**73,29**	**90,40**	87,95	**93,40**	**11,93**	**64,96**
FrALBERT-base	90,71	88,37	82,48	92,94	14,71	69,18	89,29	86,86	91,00	17,81	62,01
MEDIA 2022**, full**											
CamemBERT-base-CCNET	92,28	89,73	**85,33**	**92,87**	**11,61**	72,13	**90,86**	**88,29**	**91,06**	**14,18**	**64,14**
CamemBERT-base-Wikipedia-4GB	91,81	89,25	85,24	92,42	12,11	**72,15**	90,62	88,13	90,60	15,19	63,66
FlauBERT-oral-ft	**92,31**	**89,97**	84,26	92,34	12,68	71,54	90,23	87,76	90,68	15,11	63,10
FrALBERT-base	90,64	88,29	80,10	90,38	17,40	68,04	88,89	86,25	88,22	20,16	61,24

TABLE 3 – Meilleures performances de nos modèles optimisés par PBT sur le jeu d'évaluation de MEDIA. Les résultats ont été obtenus à partie des transcriptions manuelles (partie gauche du tableau) ou des sorties RAP (approche en cascade, partie droite).

4.1 Approche en cascade

L'approche en cascade (ou séquentielle) consiste à utiliser deux modules séparés et se suivant pour résoudre des problèmes spécifiques. Ici, un module de RAP est suivi d'un module de compréhension.

Le modèle RAP utilisé pour l'approche en cascade est constitué d'un encodeur de parole (LeBenchmark FR 3k large (Evain *et al.*, 2021)), suivi de 3 couches de bi-LSTM et d'une couche linéaire de 1024 neurones. Le modèle utilise l'optimiseur Adam avec un taux d'apprentissage de $0,0001$, tandis que la couche de sortie linéaire utilise un optimiseur Adadelta avec un taux d'apprentissage de $1,0$. La fonction de coût de la classification temporelle connexionniste (CTC) est optimisée pour 100 époques, visant le meilleur taux d'erreur de mots (WER pour *Word Error Rate* en anglais) possible. Nous obtenons un score de $9,49\%$ de WER sur MEDIA 2022 et de $10,51\%$ sur l'ensemble de données MEDIA original avec ce système.

Les sorties du module de RAP sont ensuite transmises au modèle joint présenté dans la section 3.1. Ce second modèle prédit donc les informations sémantiques (concepts et intentions) à partir des transcriptions automatiques. Les résultats du système cascade sont présentés dans la partie droite du Tableau 3.

4.2 Approche bout-en-bout

Une approche de bout-en-bout (*end-to-end* en anglais) vise à développer un système unique, directement optimisé pour extraire des informations sémantiques de la parole sans utiliser de transcriptions intermédiaires. Notre modèle de bout-en-bout est composé de l'encodeur de parole SAMU-XLSR original (Khurana *et al.*, 2022), ou SAMU-XLSR$_{IT \oplus FR}$ spécialisé (Laperrière *et al.*, 2023), suivi de deux blocs de décodage différents de 3 couches bi-LSTM de 1024 neurones. Chaque bloc est suivi d'une couche entièrement connectée de la même dimension, activée avec *Leaky ReLU* et une fonction *softmax*. Un bloc est optimisé pour produire les intentions des segments audio, tandis que l'autre exécute la tâche de détection de concepts de MEDIA.

Nous avons optimisé les fonctions de coût (*loss* en anglais) CTC sur 100 époques avec les mêmes optimiseurs que ceux utilisés dans l'approche en cascade, à l'exception de la couche linéaire de

l'optimiseur de classification des intentions dont le taux d'apprentissage est fixé à $0,1$. La somme des deux *loss* est définie comme suit : $loss = \frac{1}{4} * loss(intent) + loss(slot)$.

	Intent		Slot-filling		
Modèle	**Accuracy**	**EMR**	**F1mh**	**CER**	**SFA**
MEDIA original, relax					
SAMU-XLSR	**92,02**	**90,14**	91,68	16,01	71,57
SAMU-XLSR$_{IT \oplus FR}$	91,74	90,02	**92,35**	15,44	**72,51**
LeBenchmark FR 3k large	91,75	89,91	91,98	15,51	71,12
MEDIA 2022**, relax**					
SAMU-XLSR	**90,74**	**88,85**	90,01	15,28	68,50
SAMU-XLSR$_{IT \oplus FR}$	90,53	88,64	90,65	15,16	70,83
LeBenchmark FR 3k large	90,18	87,98	**90,77**	**15,08**	**71,12**
MEDIA 2022**, full**					
SAMU-XLSR	90,52	88,69	88,88	18,50	69,72
SAMU-XLSR$_{IT \oplus FR}$	**90,98**	**88,93**	**89,14**	**18,30**	**70,63**
LeBenchmark FR 3k large	90,07	88,02	87,99	19,67	69,06

TABLE 4 – Résultats pour l'approche bout-en-bout sur les différents corpus MEDIA.

Le Tableau 4 présente les résultats sur les tâches de classification d'intention et de détection de concepts avec cette approche de bout-en-bout. Pour la tâche de classification en intentions, l'approche de bout-en-bout obtient globalement de meilleurs résultats sur MEDIA 2022 que l'approche en cascade. Pour la tâche de détection des concepts, la tendance semble s'inverser. Cependant, les écarts entre les résultats du Tableau 4 et du Tableau 3 pourraient ne pas être suffisamment importants pour déterminer si une approche est favorable à l'autre. Néanmoins, nous obtenons de meilleurs scores de SFA avec notre approche de bout-en-bout pour toutes les versions du corpus MEDIA.

5 Conclusion

Dans cet article, nous avons présenté une version de l'ensemble de données de référence MEDIA enrichi avec des annotations de l'intention. Nous espérons ainsi ouvrir plus largement l'utilisation de ce corpus à la communauté internationale. Nous avons présenté également les premiers résultats expérimentaux sur cet ensemble de données enrichi en utilisant des modèles joints pour la classification des intentions et la détection de concepts.

Nous avons présenté différents systèmes réalisant conjointement la classification des intentions et la détection des concepts, que ce soit sur les transcriptions manuelles, les transcriptions automatiques (cascade) ou les signaux de parole (bout-en-bout). Les résultats expérimentaux sur les transcriptions manuelles et automatiques n'ont pas pu atteindre les résultats antérieurs de l'état-de-l'art pour la tâche détection de concepts, mais sont toujours compétitifs. Les modèles de bout-en-bout faisant l'optimisation jointe semblent obtenir de meilleurs résultats sur les deux tâches que les modèles en cascade.

Les annotations en intention sont librement disponibles dans un dépôt public [3] incluant le manuel d'annotation.

3. https://gitlab.lisn.upsaclay.fr/nlp/corpora/media-benchmark-intent-annotations

Références

BONNEAU-MAYNARD H., AYACHE C., BECHET F., DENIS A., KUHN A., LEFEVRE F., MOSTEFA D., QUIGNARD M., ROSSET S., SERVAN C. & VILLANEAU J. (2006). Results of the French evalda-media evaluation campaign for literal understanding. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy : European Language Resources Association (ELRA).

BONNEAU-MAYNARD H., ROSSET S., AYACHE C., KUHN A. & MOSTEFA D. (2005). Semantic annotation of the French media dialog corpus. In *Proceedings Interspeech 2005*, p. 3457–3460. DOI : 10.21437/Interspeech.2005-312.

BOULANGER H., LAVERGNE T. & ROSSET S. (2022). Generating unlabelled data for a tri-training approach in a low resourced NER task. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, p. 30–37, Hybrid : Association for Computational Linguistics. DOI : 10.18653/v1/2022.deeplo-1.4.

BÉCHET F. & RAYMOND C. (2019). Benchmarking Benchmarks : Introducing New Automatic Indicators for Benchmarking Spoken Language Understanding Corpora. In *Proceedings Interspeech 2019*, p. 4145–4149. DOI : 10.21437/Interspeech.2019-3033.

CASTELLUCCI G., BELLOMARIA V., FAVALLI A. & ROMAGNOLI R. (2019). Multi-lingual intent detection and slot filling in a joint bert-based model. *ArXiv*, **abs/1907.02884**.

CATTAN O., GHANNAY S., SERVAN C. & ROSSET S. (2022). Benchmarking Transformers-based models on French Spoken Language Understanding tasks. In *Proceedings Interspeech 2022*, p. 1238–1242. DOI : 10.21437/Interspeech.2022-385.

CATTAN O., SERVAN C. & ROSSET S. (2021). On the Usability of Transformers-based Models for a French Question-Answering Task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, p. 244–255, Held Online : INCOMA Ltd.

CHEN Q., ZHUO Z. & WANG W. (2019). Bert for joint intent classification and slot filling. *ArXiv*, **abs/1902.10909**.

CHEN Y.-N., HAKANNI-TÜR D., TUR G., CELIKYILMAZ A., GUO J. & DENG L. (2016). Syntax or semantics ? knowledge-guided joint semantic frame parsing. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, p. 348–355. DOI : 10.1109/SLT.2016.7846288.

COUCKE A., SAADE A., BALL A., BLUCHE T., CAULIER A., LEROY D., DOUMOURO C., GISSELBRECHT T., CALTAGIRONE F., LAVRIL T., PRIMET M. & DUREAU J. (2018). Snips Voice Platform : an embedded Spoken Language Understanding system for private-by-design voice interfaces. *ArXiv*, **abs/1805.10190**.

EVAIN S., NGUYEN H., LE H., BOITO M. Z., MDHAFFAR S., ALISAMIR S., TONG Z., TOMASHENKO N. A., DINARELLI M., PARCOLLET T., ALLAUZEN A., ESTÈVE Y., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2021). Lebenchmark : A reproducible framework for assessing self-supervised representation learning from speech. In *Interspeech*, p. 1439–1443.

GHANNAY S., SERVAN C. & ROSSET S. (2020). Neural networks approaches focused on French spoken language understanding : application to the MEDIA evaluation task. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 2722–2727, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.245.

GODBOLE S. & SARAWAGI S. (2004). Discriminative methods for multi-labeled classification. In H. DAI, R. SRIKANT & C. ZHANG, Éds., *Advances in Knowledge Discovery and Data Mining*, p. 22–30, Berlin, Heidelberg : Springer Berlin Heidelberg.

GOO C.-W., GAO G., HSU Y.-K., HUO C.-L., CHEN T.-C., HSU K.-W. & CHEN Y.-N. (2018). Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2018)*, p. 753–757, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-2118.

GUO D., TUR G., YIH W.-T. & ZWEIG G. (2014). Joint semantic utterance classification and slot filling with recursive neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, p. 554–559. DOI : 10.1109/SLT.2014.7078634.

HAKKANI-TÜR D., TUR 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 *Proceedings Interspeech 2016*, p. 715–719. DOI : 10.21437/Interspeech.2016-402.

HAN S. C., LONG S., LI H., WELD H. & POON J. (2021). Bi-directional joint neural networks for intent classification and slot filling. In *Interspeech*.

HEMPHILL C. T., GODFREY J. J. & DODDINGTON G. R. (1990). The ATIS Spoken Language Systems Pilot Corpus. In *Speech and Natural Language : Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

HERVÉ N., PELLOIN V., FAVRE B., DARY F., LAURENT A., MEIGNIER S. & BESACIER L. (2022). Using ASR-generated text for spoken language modeling. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, p. 17–25, virtual+Dublin : Association for Computational Linguistics. DOI : 10.18653/v1/2022.bigscience-1.2.

JADERBERG M., DALIBARD V., OSINDERO S., CZARNECKI W. M., DONAHUE J., RAZAVI A., VINYALS O., GREEN T., DUNNING I., SIMONYAN K., FERNANDO C. & KAVUKCUOGLU K. (2017). Population based training of neural networks. *ArXiv*, **abs/1711.09846**.

JEONG M. & LEE G. G. (2008). Triangular-chain conditional random fields. *IEEE Transactions on Audio, Speech, and Language Processing*, **16**(7), 1287–1302. DOI : 10.1109/TASL.2008.925143.

KHURANA S., LAURENT A. & GLASS J. (2022). Samu-xlsr : Semantically-aligned multimodal utterance-level cross-lingual speech representation. *IEEE Journal of Selected Topics in Signal Processing*, **16**(6), 1493–1504. DOI : 10.1109/JSTSP.2022.3192714.

LAN Z., CHEN M., GOODMAN S., GIMPEL K., SHARMA P. & SORICUT R. (2020). ALBERT : A Lite BERT for Self-supervised Learning of Language Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* : OpenReview.net.

LAPERRIÈRE G., NGUYEN H., GHANNAY S., JABAIAN B. & ESTÈVE Y. (2023). Semantic enrichment towards efficient speech representations. In *Interspeech*, p. 705–709.

LAPERRIÈRE G., PELLOIN V., CAUBRIÈRE A., MDHAFFAR S., CAMELIN N., GHANNAY S., JABAIAN B. & ESTÈVE Y. (2022). The Spoken Language Understanding MEDIA Benchmark Dataset in the Era of Deep Learning : data updates, training and evaluation tools. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1595–1602, Marseille, France : European Language Resources Association.

LIU B. & LANE I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. *ArXiv*, **abs/1609.01454**.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE E., SEDDAH D. & SAGOT B. (2020). CamemBERT : a Tasty French Language Model. In *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

PELLOIN V., DARY F., HERVÉ N., FAVRE B., CAMELIN N., LAURENT A. & BESACIER L. (2022). ASR-Generated Text for Language Model Pre-training Applied to Speech Tasks. In *Proceedings Interspeech 2022*, p. 3453–3457. DOI : 10.21437/Interspeech.2022-352.

QIN L., LIU T., CHE W., KANG B., ZHAO S. & LIU T. (2021). A co-interactive transformer for joint slot filling and intent detection. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* : IEEE. DOI : 10.1109/icassp39728.2021.9414110.

SOROWER M. S. (2010). A literature survey on algorithms for multi-label learning.

TANG H., JI D. & ZHOU Q. (2020). End-to-end masked graph-based crf for joint slot filling and intent detection. *Neurocomputing*, **413**, 348–359. DOI : https ://doi.org/10.1016/j.neucom.2020.06.113.

TUR G. & MORI R. D. (2011). *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*. John Wiley & Sons.

VAN ENGELEN J. E. & HOOS H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, **109**(2), 373–440. DOI : 10.1007/s10994-019-05855-6.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

WANG C., HUANG Z. & HU M. (2020). Sasgbc : Improving sequence labeling performance for joint learning of slot filling and intent detection. In *Proceedings of 2020 6th International Conference on Computing and Data Engineering*, ICCDE '20, p. 29–33, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3379247.3379266.

WELD H., HUANG X., LONG S., POON J. & HAN S. C. (2022). A survey of joint intent detection and slot filling models in natural language understanding. *ACM Comput. Surv.*, **55**(8). DOI : 10.1145/3547138.

XU P. & SARIKAYA R. (2013). Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, p. 78–83. DOI : 10.1109/ASRU.2013.6707709.

XU W., HAIDER B. & MANSOUR S. (2020). End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 5052–5063, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.410.

ZHOU Z.-H. & LI M. (2005). Tri-training : exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, **17**(11), 1529–1541. DOI : 10.1109/TKDE.2005.186.

ZHUANG L., WAYNE L., YA S. & JUN Z. (2021). A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, p. 1218–1227, Huhhot, China : Chinese Information Processing Society of China.