Le benchmark MEDIA revisité : données, outils et évaluation dans un contexte d'apprentissage profond

Gaëlle Laperrière[1] Valentin Pelloin[2] Antoine Caubrière[1] Salima Mdhaffar[1]
Nathalie Camelin[2] Sahar Ghannay[3] Bassam Jabaian[1] Yannick Estève[1]
(1) LIA - Avignon Université, France
(2) LIUM - Le Mans Université, adresse, France
(3) Université Paris-Saclay, CNRS, LISN, 91405 Orsay, France
[1]prénom.nom@univ-avignon.fr, [2]prénom.nom@univ-lemans.fr,
[3]prénom.nom@limsi.fr

RÉSUMÉ

Nous discutons ici du jeu de données françaises de référence MEDIA, créé en 2005 et distribué par ELRA gratuitement pour la recherche académique depuis 2020. Bien que parmi les plus riches et complexes à traiter, ces données sont rarement utilisées au-delà de la communauté scientifique française. Pour en faciliter l'usage dans un contexte d'apprentissage profond, une recette complète a été intégrée à SpeechBrain, une boite à outils logicielle dédiée au traitement de la parole par des approches neuronales, de plus en plus populaire au niveau international. De plus, des corrections ont été apportées aux annotations manuelles, proposées par différents chercheurs ayant régulièrement travaillé sur ces données. Cette nouvelle version du corpus sera intégrée au catalogue de ELRA. Un nouvel ensemble de données jamais utilisées jusqu'à présent, mais collectées durant la création du corpus original, est également décrit. Enfin, nous abordons des considérations liées à l'évaluation de la tâche MEDIA.

ABSTRACT

The Spoken Language Understanding MEDIA Benchmark Dataset in the Era of Deep Learning : data updates, training and evaluation tools

With the emergence of neural end-to-end approaches for spoken language understanding, a growing number of studies on speech intent detection have been presented during these last three years and new benchmark datasets have been produced. In this paper, we focus on the French benchmark dataset MEDIA, created in 2005 and distributed freely for academic research by ELRA since 2020. This dataset has been shown as being the most challenging one in its domain but is unfortunately not much used beyond the French research community. To facilitate its use, a complete recipe has been integrated to SpeechBrain, a toolkit based on PyTorch. Corrections were brought to the initial manual annotations. The new version of MEDIA will be distributed by ELRA. We used a significant amount of data collected during the construction of MEDIA and never used until now to create a new corpus called test2. Last, we discuss evaluation issues.

MOTS-CLÉS : Compréhension de la parole, Corpus MEDIA, Modèles neuronaux de bout-en-bout .

KEYWORDS: Spoken Language Understanding, MEDIA dataset, End-to-end neural networks .

Cet article est partiellement financé par la Commission Européenne via le projet SELMA, contrat 957017, et par le projet AISSPER porté par l'Agence Nationale de la Recherche (ANR) sous le contrat ANR-19-CE23-0004-01.

1 Introduction

La compréhension de la parole (SLU) fait référence aux tâches de traitement automatique du langage naturel liées à l'extraction d'informations sémantiques depuis le signal de la parole. Elle est considérée tel "*un domaine à l'intersection du traitement de la parole et du langage naturel, tirant parti des technologies d'apprentissage automatique et d'intelligence artificielle*" (Tur & De Mori, 2011). En interaction humain-machine, elle vise à convertir une entrée utilisateur en représentation sémantique. Les trois points primordiaux en SLU sont : 1) la représentation sémantique du domaine en lien avec l'application visée 2) l'extraction de la parole et la projection de la sémantique vers la cette représentation 3) l'évaluation du système réalisé.

La représentation sémantique dépend de la tâche. Elle est généralement construite via des concepts sémantiques portés par des séquences de mots. Pour le corpus ATIS (Air Travel Information System) (Hemphill *et al.*, 1990), les étiquettes sémantiques sont dédiées à la tâche de planification de vols. Sa représentation sémantique comporte 17 intentions (avion, distance, vol...) et des concepts sémantiques (départ_date.relative...), associés à chaque trame. D'autres corpus SLU existent (Coucke *et al.*, 2018; Shah *et al.*, 2018; Lugosch *et al.*, 2021), avec leur propre représentation sémantique, souvent basée sur une liste d'étiquettes spécifiques.

Ces deux dernières décennies, les principales approches pour la tâche SLU utilisaient des algorithmes d'apprentissage automatique, génératifs et discriminants (Raymond & Riccardi, 2007). Jusqu'au début des années 2010, les approches les plus populaires pour l'étiquetage de mots étaient les champs aléatoires conditionnels (CRF) (Hahn *et al.*, 2010). Depuis l'émergence de l'apprentissage profond, les architectures neuronales mènent le pas, avec des réseaux récurrents (Mesnil *et al.*, 2013; Kurata *et al.*, 2016; Dupont *et al.*, 2017) et encodeurs-décodeurs avec mécanismes d'attention (Simonnet *et al.*, 2017; Li *et al.*, 2018). Les derniers travaux utilisent des modèles de représentation du langage pré-appris de manière auto-supervisée, comme BERT (Devlin *et al.*, 2019), améliorant les résultats de l'état-de-l'art dans différentes tâches du domaine (Korpusik *et al.*, 2019; Ghannay *et al.*, 2020). Ces modèles utilisent une approche cascade : un système de reconnaissance de la parole (ASR) transcrit automatiquement l'énoncé de l'utilisateur, puis un système de compréhension du langage naturel (NLU) en extrait la sémantique. La transcription intermédiaire peut contenir des erreurs, propagées au système NLU. Les approches de bout-en-bout ont donc été proposées pour pallier cette propagation d'erreurs en n'utilisant pas de transcription intermédiaire. Elles permettent d'optimiser l'entièreté du modèle pour la tâche finale d'extraction de sémantique, tandis que les modèles cascade nécessitent une optimisation indépendante de chaque sous-tâche (ASR puis NLU). Les systèmes de bout-en-bout peuvent générer une transcription, mais aussi des concepts sémantiques (Ghannay *et al.*, 2018; Desot *et al.*, 2019; Dinarelli *et al.*, 2020; Evain *et al.*, 2021).

L'évaluation de la prédiction dépend de la complexité de la représentation sémantique. Pour la détection d'intention, la précision, le rappel et le taux de classification correct sont généralement suffisants. Dans des cas plus complexes comme l'étiquetage sémantique, il faut au moins considérer l'étiquette sémantique et sa valeur. Il est possible d'utiliser le taux d'erreur de concepts (CER) et le taux d'erreur concepts-valeurs (CVER), souvent utilisé dans le cadre du corpus MEDIA.

Dans cet article, nous décrivons le jeu de données françaises MEDIA dédié à la tâche SLU (Béchet & Raymond, 2019). Bien que parmi les plus riches et complexes, il est rarement utilisé au-delà de la communauté scientifique française. Pour faciliter son utilisation et rendre les données MEDIA plus accessibles, nous présentons une recette persistante avec préparation des données, apprentissage de bout-en-bout et évaluations, intégrée à SpeechBrain, une boite à outil tout-en-un, *open-source*, et dûment maintenue. Des corrections ont été apportées aux annotations manuelles de MEDIA et de

nombreuses données collectées, mais jamais utilisées ont été regroupées pour former un nouveau corpus de test. Nous présentons ses résultats avec la recette SpeechBrain.

2 Jeu de données MEDIA original

Le benchmark français MEDIA (Bonneau-Maynard *et al.*, 2005) a été crée dans le cadre du projet Technolangue du gouvernement français en 2002. Il est dédié à l'extraction de l'information sémantique à partir de la parole dans un contexte de dialogues humain-machine pour une tâche de réservation de chambre d'hôtel. Il vise entre autres à mettre en place une infrastructure de production et de diffusion de ressources linguistiques, et d'évaluation des technologies de la langue écrite et orale. Le corpus est distribué par ELRA gratuitement pour la recherche académique depuis 2020 .

2.1 Données

Le corpus MEDIA est composé d'enregistrements de dialogues téléphoniques avec leurs transcriptions manuelles et leurs annotations sémantiques. Il compte 1258 dialogues, de 250 locuteurs différents. Il a été enregistré à l'aide d'une méthode de type magicien d'Oz (WoZ) (Green & Wei-Haas, 1985; Dahlbäck *et al.*, 1993) : un humain (le "Wizard") prétend être un ordinateur, tandis que l'utilisateur est amené à croire qu'il interagit avec une machine intelligente. Dans MEDIA, seule la transcription du dialogue utilisateur est enrichie d'annotations sémantiques.

Les tables 1, 2 montrent qu'une partie conséquente des données MEDIA n'a pas été utilisée lors de la campagne officielle en 2005, car finalisées après la fin de la campagne. Par conséquent, même si ces données sont présentes dans l'archive distribuée par ELRA, elles ne sont pas répertoriées officiellement et sont cachées parmi les sous-répertoires qui structurent l'archive MEDIA. À notre connaissance, ces données n'ont jamais été utilisées dans des travaux de recherche jusqu'à présent.

Données	**Nb. Échantillons**	**Nb. Tours de parole**	**Nb. Dialogues**
train	13,7 k	13,0 k	727
dev	1,4 k	1,3 k	79
test	3,8 k	3,5 k	208
pas utilisées	4,0 k	3,8 k	244

TABLE 1 – Données MEDIA d'origine en ne tenant compte que des énoncés de l'utilisateur.

Données	**Temps des échantillons**			**Temps globaux**		
	Nb. Heures	**Durée Moyenne**	**Durée Médiane**	**Nb. Heures**	**Durée Moyenne**	**Durée Médiane**
train	16h56m	4,69s	3,12s	42h10m	209s	194s
dev	01h40m	4,77s	2,79s	03h37m	165s	158s
test	04h47m	4,89s	3,34s	11h34m	200s	190s
pas utilisées	05h35m	5,30s	3,86s	14h30m	214s	196s

TABLE 2 – Statistiques sur la durée des échantillons de l'utilisateur avant segmentation et sur les temps globaux des enregistrements audio (utilisateur, WoZ et blancs de paroles compris).

`http://catalog.elra.info/en-us/repository/browse/ELRA-E0024/`
International Standard Language Resource Number : 699-856-029-354-6

Le dictionnaire sémantique MEDIA comprend 83 attributs dont 73 basiques comme *localisation*, 4 modificateurs comme *distance-relative* et 6 attributs généraux tels que *attribut-connecteur*, en plus de 19 spécificateurs (Bonneau-Maynard *et al.*, 2006) comme *début, fin*, spécialisant leurs rôle dans le contexte. Des phénomènes linguistiques complexes, comme les co-références, sont également gérés. La combinaison des attributs et des spécificateurs donne 1121 concepts possibles. Ces attributs sont supportés par des mots ou séquences de mots appelés **mots-support**. Pour chaque attribut présent dans l'annotation sémantique, deux autres informations sont fournies : son mode et sa valeur normalisée. Quatre modes existent : '+' affirmatif, '-' négatif, ' ?' interrogatif ou '~' optionnel. Un énoncé typique du benchmark MEDIA annoté par séquences de quadruplets (mots-support, mode, attribut, valeur normalisée) serait : (je veux réserver, +, *réservation*, `réservation`) (une, +, *nombre-chambre*, `1`) (chambre double, +, *chambre-type*, `chambre double`) (jusqu'à, +, *comparative-paiement*, `moins de`) (cent, +, *paiement-montant*, `100`) (euros, +, *paiement-monnaie*, `euro`).

L'étude de (Béchet & Raymond, 2019) a mis en évidence le fait que la tâche MEDIA peut être considérée comme le benchmark SLU le plus riche et complexe à traiter, par rapport à d'autres benchmarks bien connus tels que ATIS (Dahl *et al.*, 1994), SNIPS (Coucke *et al.*, 2018), et M2M (Shah *et al.*, 2018).

2.2 Évaluation

La métrique d'évaluation utilisée pendant la campagne d'évaluation 2005 (Bonneau-Maynard *et al.*, 2006) est le "*understanding error rate*", soit taux d'erreur de compréhension. Elle aligne la représentation sémantique de référence à l'hypothèse avec la distance de Levenshtein, puis compte le nombre d'insertions, suppressions et substitutions de triplets comme présentés précédemment. Cette métrique est similaire au taux d'erreur de mots (*word error rate*), utilisé en ASR.

Deux notations sont utilisées pour les concepts sémantiques de MEDIA : la notation *Full* considère tous les attributs possibles (1121 possibilités) tandis que la notation *Relax* ne considère pas les spécificateurs (83 possibilités). Le mode peut aussi être réduit au choix binaire 'négatif' et 'affirmatif'.

2.3 Problèmes rencontrés

Évolution des métriques Après la campagne de 2005, nous n'avons compté qu'une étude poursuivant l'évaluation en *taux d'erreur de compréhension*, avec les différentes notations et modes (Lehuen & Lemeunier, 2010). Le *taux d'erreur de concepts* (CER) a été introduit par (Raymond & Riccardi, 2007) pour simplifier l'évaluation du benchmark MEDIA est devenu *de facto* la métrique de référence (Hahn *et al.*, 2010; Dinarelli *et al.*, 2020; Ghannay *et al.*, 2018) dans les travaux récents sur MEDIA. Il fonctionne comme le taux d'erreur de compréhension, mais ne compte que les alignements des étiquettes de mots nommées **concepts**. L'alignement des concepts et de leur valeur normalisée a été utilisé par (Hahn *et al.*, 2010), puis nommé *taux d'erreur concepts-valeurs* (CVER) par (Simonnet *et al.*, 2017, 2018). En complément du CER, cette métrique a été utilisée dans de nombreux travaux (Caubrière *et al.*, 2019; Ghannay *et al.*, 2021; Pelloin *et al.*, 2021), suite à (Simonnet *et al.*, 2017) et (Simonnet *et al.*, 2018).

Normalisation de valeur pour CVER (Hahn *et al.*, 2010) a conclu que normaliser les valeurs en utilisant des règles manuellement établies grâce au corpus d'apprentissage était une solution des plus performantes. Elles permettent de mieux normaliser les dates et nombres, aux vues de la quantité limitée de données. Beaucoup des études récentes sur le benchmark MEDIA utilisent ces règles.

(Pelloin *et al.*, 2021) a proposé la normalisation automatique directe des valeurs après prédiction

grâce à un encodeur-décodeur de bout-en-bout avec mécanismes d'attention. De très bons résultats ont été obtenus, mais comme pour d'autres études (Hahn *et al.*, 2010), les règles établies restent plus performantes.

Afin de donner des résultats justes et indépendants d'une normalisation de valeur qui pourrait être biaisée ou inefficace, nous avons évalué nos systèmes avec un CVER non-normalisé (*u-CVER*). En effet, le CVER avec valeurs normalisées par règle tel qu'utilisé dans les travaux de la dernière décennie présente des défauts majeurs. Sur les transcriptions de référence, le CVER devrait être égal à 0%. Or, en appliquant les même règles de normalisation que dans les travaux cités plus haut, nous obtenons 4,7% de CVER sur le corpus de développement, et 5,7% sur le corpus de test. Toutes ces erreurs sont des substitutions, donc des mots-support pas ou mal normalisés. Ceci justifie l'utilisation du *u-CVER*, plus strict, mais moins biaisé.

Erreurs d'annotation manuelles Le projet MEDIA a établi un schéma complet d'annotations sémantiques pour des annotateurs humains. Le langage naturel étant sujet à interprétation, des erreurs ont pu être faites et induire de fausses erreurs d'évaluation. Nous en proposons la correction. Nous traitons aussi des problèmes audio, comme la présence de tonalité de fin d'appel dans les segments.

Préparation des données Les annotations des données MEDIA considèrent des détails de prononciation, comme des mots tronqués ou trop proches, indiqués par des parenthèses ou astérisques. Ces annotations peuvent ne pas avoir été traitées de la même manière dans les différentes études, rendant les résultats expérimentaux publiés difficilement comparables. Nous fournissons donc nos scripts de préparation des données MEDIA dans notre recette SpeechBrain.

Intégration des données non-utilisées Un nombre important de données ont été collectées, mais jamais utilisées (*cf. Section 2.1*). Nous les avons intégrées à notre recette et partageons ici leurs premiers résultats expérimentaux.

3 Révision des données

Des corrections ont été apportées aux annotations manuelles de MEDIA. Cette nouvelle version sera distribuée prochainement par ELRA. Une normalisation a été réalisée sur la transcription et les concepts. Outre leur orthographe, nous avons retiré les redondances d'espaces, corrigé des connexions d'apostrophes et traits d'union, et la casse de certains noms propres. Le canal audio (droite ou gauche) d'enregistrement vocal de l'utilisateur est maintenant indiqué. L'identifiant de certains utilisateurs a été corrigé pour en respecter le format.

Corpus	Occurrences			Lexique			
	Mots	**Mots Tronqués**	**Concepts Full et Relax**	**Mots**	**Mots Tronqués**	**Concepts Full**	**Concepts Relax**
train	92,6 k	820	31,7 k	2,3 k	372	144	73
dev	10,5 k	134	3,3 k	0,8 k	89	104	63
test	26,0 k	227	8,8 k	1,4 k	146	125	71
test2	28,0 k	159	9,4 k	1,3 k	107	129	71

TABLE 3 – Nombre d'occurrences et taille du lexique de mots, mots tronqués et concepts en notation Full et Relax dans la nouvelle version de MEDIA, pour l'utilisateur uniquement.

De nombreuses données avaient été générées, mais jamais utilisées. Nous avons décidé de les utiliser

pour créer le corpus *test2*, similaire à *test*, détaillé dans les tables 1, 2. Nous avons dû générer la notation Relax des concepts sémantiques présents dans test2, car les données étaient uniquement annotées en notation Full.

Nous présentons les nouvelles statistiques des données MEDIA avec le nouveau corpus *test2*. La table 3 donne les statistiques de mots et mots tronqués dans la transcription de MEDIA (*cf. Section 2.3*) et celles des concepts, dont le nombre d'occurrences est le même en notation Full et Relax.

4 Recette SpeechBrain MEDIA

Une recette complète a été intégrée à SpeechBrain, une boite à outils très utilisée, tout-en-un et open-source pour des tâches d'Intelligence Artificielle conversationnelle. Cet outil propose de nombreuses autres recettes, prêtes à l'emploi. L'avantage d'intégrer notre recette MEDIA dans SpeechBrain est la garantie de garder un code source persistant et maintenu au gré des évolutions des prochaines années.

La recette comprend les scripts de traitement des données MEDIA, d'apprentissage et d'évaluation pour la tâche ASR (sans concepts sémantiques) et la tâche SLU. Nous les détaillons dans la prochaine partie. Elle utilise une architecture de bout-en-bout permettant de réaliser le *fine-tuning* d'un modèle wav2vec 2.0 (Baevski *et al.*, 2020). Les modèles wav2vec 2.0 sont pré-appris en générant une représentation de la parole de manière auto-supervisée, avec d'importantes quantités de données. Ils utilisent des couches neuronales convolutionnelles et un Transformer. Il est possible de les *fine-tuner* via un apprentissage supervisé, comme fait avec les données MEDIA.

4.1 Préparation des données

Au lancement de la recette, les données MEDIA sont préparées automatiquement. Les paramètres d'apprentissage peuvent être optimisés, dont le tri en durée croissante des échantillons, plus efficace.

La majorité des caractères spéciaux ont été retirés. Nous avons gardé les chevrons servant à encadrer les concepts et leurs mots-support (" <heure-départ> midi >"). L'apostrophe a été gardé et rattaché au mot précédent pour limiter la taille du vocabulaire, hormis pour "c'est" car très commun. Seuls les traits d'union des nombres ont été retirés pour la même raison car ils n'aident pas à comprendre le dialogue. Des astérisques ont été ajoutés aux mots tronqués, changeant "bon(jour)", une prononciation ambiguë de "bonjour", par "bon*". L'indication est donc gardée sans créer un nouveau mot dans le lexique.

Nous avons utilisé les balises de synchronisation temporelels des fichiers xml originels de MEDIA pour segmenter des échantillons. Elles permettent aussi de retirer des échantillons les blancs ou tonalités de fin d'appel dans le signal. La table 4 donne les nouvelles durées des échantillons utilisateur après segmentation.

4.2 Architecture Neuronale et Apprentissage

La recette utilise le modèle LeBenchmark (Evain *et al.*, 2021) wav2vec2-FR-3k large, pré-appris sur 3k heures de parole en français. Les modèles proposés par LeBenchmark sont librement accessibles. Au dessus de la couche la plus haute du wav2vec 2.0, nous avons ajouté 3 couches denses de 512 neurones, activées par LeakyReLU, puis une couche linéaire de même dimension et une couche Softmax. La fonction de coût utilisée est la fonction Connectionist Temporal Classification (Graves *et al.*, 2006).

https://github.com/speechbrain/speechbrain/tree/develop/recipes/MEDIA

Corpus	Nb. Heures	Durée Moyenne	Durée Médiane
train	10h52m	2,85s	1,69s
dev	01h13m	3,23s	1,91s
test	03h01m	2,88s	1,70s
test2	03h16m	2,94s	1,93s
total	18h22m	2,90s	1,75s

TABLE 4 – Durée des échantillons utilisateur après segmentation et préparation des données.

Les poids de ces couches ajoutées sont initialisés aléatoirement et ceux du wav2vec 2.0 sont ceux calculés durant son pré-apprentissage.

Le réseau neuronal reçoit en entrée le signal audio échantillonné à 16kHz issu de fichiers wav, et génère les caractères de l'alphabet nécessaire à la tâche MEDIA. Un décodeur glouton à la sortie de la couche Softmax sélectionne les caractères finaux que nous évaluons.

L'apprentissage supervisée SLU MEDIA est réalisé avec un *fine-tuning* de notre wav2vec 2.0. La recette propose une seconde solution, utilisant un wav2vec 2.0 fine-tuné sur 425,5 heures de données françaises issus CommonVoice (version 6.1) pour une tâche de reconnaissance de la parole. En ré-initialisant les poids de sa dernière couche, nous réalisons le même apprentissage que pour la première solution. Dans les prochaines parties, nous présenterons les résultats de ces deux architectures :

— **media-base** avec le wav2vec 2.0 *fine-tuné* directement sur MEDIA pour la tâche SLU.
— **media-comvoice** avec le wav2vec 2.0 *fine-tuné* sur CommonVoice pour une tâche ASR, puis sur MEDIA pour la tâche SLU.

Nous utilisons l'optimiseur Adam pour le wav2vec 2.0, pour un taux d'apprentissage de $0,0001$, et AdaDelta pour les autres couches, avec un taux de 1 et momentum de $0,95$, le tout sur 30 époques.

4.3 Résultats préliminaires

Notre recette propose trois métriques :

— Le taux d'erreur de caractères **ChER**. Chaque caractère est une unité à évaluer (un concept est codé par un unique caractère).
— Le taux d'erreur de concepts **CER**. Seuls les concepts prédits, comptés comme une unité, sont comparés à la référence.
— Le taux d'erreur de concepts et valeurs non-normalisées **u-CVER**. L'ensemble concept et valeur forme une unique unité. Un caractère erroné de la valeur prédite rend l'entière unité fausse.

Bien que les concepts MEDIA aient été en notation Relax dans les dernières publications, nous décidons de ré-introduire la notation Full car elle amène un plus grand défi.

Un modèle différent, donc un lexique sémantique différent, a été appris pour chaque notation. La table 5 présente les résultats des modèles **media-base** et **media-comvoice** (*cf. Section4.2*). Nos résultats sont légèrement sous ceux de l'état de l'art de bout-en-bout, avec 16.3% de CER sur le test. Cela prouve la validité de la recette qui nécessite une simple optimisation de ses paramètres.

La table 6 présente les résultats du meilleur système **media-comvoice** sur le corpus test2. Étant similaires à ceux du corpus test pour le même modèle, le corpus test2 est donc bien adapté à notre

https ://commonvoice.mozilla.org/fr/datasets

tâche. Le corpus MEDIA, bien qu'utilisé depuis une quinzaine d'années, ne semble pas avoir subi de sur-ajustement par ses données de test, volontairement ou non.

Notation	**Modèle**	**dev**			**test**		
		ChER	CER	u-CVER	ChER	CER	u-CVER
Full	**media-base**	8,4	28,9	41,2	8,2	26,1	37,5
	media-comvoice	7,2	24,0	34,4	6,9	20,3	30,8
Relax	**media-base**	8,1	23,3	37,1	7,9	21,8	34,1
	media-comvoice	6,8	18,1	30,4	6,7	16,3	27,7

TABLE 5 – Résultats sur les données préparées de MEDIA, pour les modèles **media-base** et **media-comvoice** et les notations Full et Relax.

Notation	**test2**		
	ChER	CER	u-CVER
Full	6,7	21,1	30,9
Relax	6,4	16,4	27,1

TABLE 6 – Résultats sur le test2 avec le modèle **media-comvoice** et les notations Full et Relax.

Conclusion

Dans cet article, ont été présentées les corrections apportées au benchmark MEDIA pour la tâche SLU, prochainement distribuées par ELRA, et la création d'un nouveau corpus de test grâce à des données collectées, mais jamais utilisées jusqu'à présent. Nous visons à faciliter l'utilisation de ce jeu de données gratuit pour la recherche scientifique. Nous proposons aussi une recette complète, intégrée à SpeechBrain, un outil déjà très populaire, open-source et tout-en-un.

Nous espérons faire grandir la communauté MEDIA en rendant accessible notre recette, qui peut être modifiée très facilement en raison de la modularité de la boite à outils SpeechBrain. MEDIA reste un des benchmarks les plus riches et complexes à traiter, y compris dans le domaine de l'apprentissage profond, et peut permettre de réelles innovations en terme d'interactions humain-machine dans de nombreuses domaines.

Références

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. F. BALCAN & H. LIN, Eds., *Advances in Neural Information Processing Systems*, volume 33, p. 12449–12460 : Curran Associates, Inc.

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

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 *INTERSPEECH*.

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. In *20th Annual Conference of the International Speech Communication Association (InterSpeech)*, p. 1198–1202.

COUCKE A., SAADE A., BALL A., BLUCHE T., CAULIER A., LEROY D., DOUMOURO C., GISSELBRECHT T., CALTAGIRONE F., LAVRIL T. *et al.* (2018). Snips voice platform : an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv :1805.10190*.

DAHL D. A., BATES M., BROWN M. K., FISHER W. M., HUNICKE-SMITH K., PALLETT D. S., PAO C., RUDNICKY A. & SHRIBERG E. (1994). Expanding the scope of the atis task : The atis-3 corpus. In *HUMAN LANGUAGE TECHNOLOGY : Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

DAHLBÄCK N., JÖNSSON A. & AHRENBERG L. (1993). Wizard of oz studies—why and how. *Knowledge-based systems*, **6**(4), 258–266.

DESOT T., PORTET F. & VACHER M. (2019). Towards end-to-end spoken intent recognition in smart home. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, p. 1–8 : IEEE.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, Minneapolis, Minnesota : Association for Computational Linguistics.

DINARELLI M., KAPOOR N., JABAIAN B. & BESACIER L. (2020). A data efficient end-to-end spoken language understanding architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 8519–8523 : IEEE.

DUPONT Y., DINARELLI M. & TELLIER I. (2017). Label-dependencies aware recurrent neural networks. In *International Conference on Computational Linguistics and Intelligent Text Processing*, p. 44–66 : Springer.

EVAIN S., NGUYEN H., LE H., BOITO M. Z., MDHAFFAR S., ALISAMIR S., TONG Z., TOMASHENKO N., DINARELLI M., PARCOLLET T. *et al.* (2021). Task agnostic and task specific self-supervised learning from speech with lebenchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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 *2018 IEEE Spoken Language Technology Workshop (SLT)*, p. 692–699 : IEEE.

GHANNAY S., CAUBRIÈRE A., MDHAFFAR S., LAPERRIÈRE G., JABAIAN B. & ESTÈVE Y. (2021). Where are we in semantic concept extraction for spoken language understanding ? In *International Conference on Speech and Computer*, p. 202–213 : Springer.

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.

GRAVES A., FERNÁNDEZ S., GOMEZ F. & SCHMIDHUBER J. (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, p. 369–376.

GREEN P. & WEI-HAAS L. (1985). The rapid development of user interfaces : Experience with the wizard of oz method. *Proceedings of the Human Factors Society Annual Meeting*, **29**(5), 470–474.

HAHN S., DINARELLI M., RAYMOND C., LEFEVRE F., LEHNEN P., DE MORI R., MOSCHITTI A., NEY H. & RICCARDI G. (2010). Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(6), 1569–1583.

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*.

KORPUSIK M., LIU Z. & GLASS J. (2019). A comparison of deep learning methods for language understanding. In *Interspeech, September 15–19, 2019, Graz, Austria*, Graz, Austria.

KURATA G., XIANG B., ZHOU B. & YU M. (2016). Leveraging sentence-level information with encoder lstm for semantic slot filling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2077–2083.

LEHUEN J. & LEMEUNIER T. (2010). A robust semantic parser designed for spoken dialog sytems. *2010 IEEE Fourth International Conference on Semantic Computing*, p. 52–55.

LI C., LI L. & QI J. (2018). A self-attentive model with gate mechanism for spoken language understanding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3824–3833.

LUGOSCH L., PAPREJA P., RAVANELLI M., HEBA A. & PARCOLLET T. (2021). Timers and such : A practical benchmark for spoken language understanding with numbers. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

MESNIL G., HE X., DENG L. & BENGIO Y. (2013). Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech*, p. 3771–3775.

PELLOIN V., CAMELIN N., LAURENT A., DE MORI R., CAUBRIÈRE A., ESTÈVE Y. & MEIGNIER S. (2021). End2end acoustic to semantic transduction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 7448–7452.

RAYMOND C. & RICCARDI G. (2007). Generative and Discriminative Algorithms for Spoken Language Understanding. In *Interspeech 2007 - 8th Annual Conference of the International Speech Communication Association*, Anvers, Belgium.

SHAH P., HAKKANI-TÜR D., TÜR G., RASTOGI A., BAPNA A., NAYAK N. & HECK L. (2018). Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv :1801.04871*.

SIMONNET E., GHANNAY S., CAMELIN N. & ESTÈVE Y. (2018). Simulating ASR errors for training SLU systems. In *LREC 2018*, Miyazaki, Japan.

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

TUR G. & DE MORI R. (2011). *Spoken language understanding : Systems for extracting semantic information from speech*. John Wiley & Sons.