Estimation de la qualité d'un système de reconnaissance de la parole pour une tâche de compréhension

Olivier Galibert [1] Nathalie Camelin [2] Paul Deléglise [2] Sophie Rosset [3]
(1) LNE, F-78190 Trappes, France
(2) LIUM - Université du Maine, Le Mans, France
(3) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
olivier.galibert@lne.fr, nathalie.camelin@lium.univ-lemans.fr,
paul.deleglise@lium.univ-lemans.fr, sophierosset@limsi.fr

RÉSUMÉ

Nous nous intéressons à l'évaluation de la qualité des systèmes de reconnaissance de la parole étant donné une tâche de compréhension. L'objectif de ce travail est de fournir un outil permettant la sélection d'un système de reconnaissance automatique de la parole le plus adapté pour un système de dialogue donné. Nous comparons ici différentes métriques, notamment le WER, NE-WER et ATENE métrique proposée récemment pour l'évaluation des systèmes de reconnaissance de la parole étant donné une tâche de reconnaissance d'entités nommées. Cette dernière métrique montrait une meilleure corrélation avec les résultats de la tâche globale que toutes les autres métriques testées. Nos mesures indiquent une très forte corrélation avec la mesure ATENE et une moins forte avec le WER.

ABSTRACT

Quality estimation of a Speech Recognition System for a Spoken Language Understanding task.

In this papier, we are interested in evaluating the quality of speech recognition system outputs considering a spoken language understanding task. The objective is to provide a tool that can select the most suitable automatic speech recognition system for a given dialogue system. We use different metrics in this study, including the WER, NE-WER and, ATENE. The latter metric was recently proposed for evaluating speech recognition systems considering a named entity recognition task. ATENE showed the better correlation with the results of the overall task among all other tested metrics. In this paper, the results indicate a stronger correlation between ATENE and the standard evaluation measure of spoken the language understanding task than with the other metrics.

MOTS-CLÉS : reconnaissance de la parole, compréhension, métrique d'évaluation.

KEYWORDS: Automatic Speech Recognition, Spoken Language Understanding, Evaluation Metric .

1 Introduction

Pouvoir estimer la qualité d'un système de Reconnaissance Automatique de la Parole (RAP) peut présenter un intérêt certain, notamment si l'on n'a pas accès au développement du système de RAP lui-même.

Actuellement, lorsqu'on mesure la qualité d'un système de RAP on utilise le plus souvent le taux

d'erreur mots (ou WER pour *Word Error Rate* (Pallett, 2003)). Or plusieurs travaux ont montré que le WER ne présentait pas toujours une bonne corrélation avec le score d'une tâche plus globale incluant la RAP. Par exemple si dans (Munteanu *et al.*, 2006), les auteurs observent, sur l'utilisabilité par des humains d'archives Web transcrites automatiquement, que la relation est linéaire entre le WER et la performance des humains à rechercher et trouver des informations dans les documents, ils observent aussi qu'en cas de grande différence de WER, celui-ci reste une métrique fiable. (Przybocki *et al.*, 1999) ont observé une forte corrélation entre le WER et la mesure utilisée pour l'évaluation de la détection des entités nommées (le SER, *Slot Error Rate*). Cette corrélation n'est pas du tout retrouvée dans une tâche similaire (voir (Ben Jannet *et al.*, 2015a)). Ceci dit, dans cette expérience, (Przybocki *et al.*, 1999) observent également que le système de RAP permettant d'aboutir au meilleur score global est celui qui obtient seulement la cinquième position en terme de WER.

Dans le contexte de la compréhension de la parole (ou SLU pour *Spoken Language Understanding*), des travaux (voir (Riccardi & Gorin, 1998; Wang *et al.*, 2003)) ont mis en évidence le fait que le WER n'est pas la métrique idéale pour évaluer la transcription pour la compréhension automatique de la parole. D'autres auteurs ont signalé des observations similaires sur les tâches de productions par des humains de résumés de séminaires préalablement transcrits automatiquement (Favre *et al.*, 2013). Dans le contexte des entités nommées, (Ben Jannet *et al.*, 2015a) ont proposé la métrique ATENE qui offre une meilleure corrélation que le WER par rapport au résultat global obtenu par des systèmes de Reconnaissance d'Entités Nommées (REN).

L'objectif de notre travail est, dans un premier temps, de voir si les résultats obtenus par ATENE sur une tâche de REN sont reproductibles sur une tâche de compréhension de parole ; dans un deuxième temps, nous souhaitons voir à quel point il est nécessaire ou pas de développer des modèles spécifiques étant donné que ces deux tâches (REN et SLU) sont conceptuellement proches.

L'article est organisé de la façon suivante : la section 2 présente quelques unes des métriques utilisées ou ayant été proposées pour l'évaluation de la qualité des systèmes de RAP. La section 3 présente le contexte de nos travaux et en particulier la tâche et les données utilisées. Puis la section 4.2 présente les expériences que nous avons menées pour répondre à nos deux questions ainsi que les différents résultats. Enfin la section 5 conclut ces travaux tout en ouvrant quelques perspectives.

2 Métriques d'évaluation utilisées en RAP

La métrique la plus utilisée pour évaluer la qualité d'un système de RAP est le WER. Cette métrique consiste à compter les erreurs selon les types prédéfinis que sont l'insertion, la suppression et la substitution déterminés par un alignement de Levenshtein entre la transcription manuelle (référence) et la transcription automatique (hypothèse). Plusieurs métriques alternatives existent. Certaines tentent de mesurer une perte d'information générale comme le RIL, *Relative Information Loss*, proposé par (Miller, 1955). Le RIL est fondé sur le principe d'information mutuelle et permet d'obtenir une mesure de la dépendance statistique entre le vocabulaire de la référence et celui de l'hypothèse. Cette mesure est représentée en termes d'entropie de Shannon. Par la suite, le WIL (Morris *et al.*, 2004) (*Word Information Loss*), qui est une approximation du RIL, a été introduit. Pour les taux d'erreurs élevés, le RIL et le WIL montrent des résultats intéressants (Morris *et al.*, 2004; McCowan *et al.*, 2004). Toujours dans le but de mesurer la perte d'information, (McCowan *et al.*, 2004) ont proposé d'adapter les métriques standards utilisées en extraction d'information, la précision (P), le rappel (R) et la f-mesure (F). L'idée général consiste à calculer le rappel et la précision au niveau des mots

en s'appuyant sur l'alignement entre la référence et l'hypothèse tel qu'il est produit par le calcul du WER. Deux autres métriques ont été proposées spécifiquement pour permettre d'évaluer la perte d'information étant donnée une tâche de reconnaissance d'entités nommées. La première métrique est le NE-WER (*Named Entity Word Error Rate*) proposé par (Garofolo *et al.*, 1999), qui mesure un WER dans les zones où la référence comprend une entité nommée. La seconde est ATENE (*Automatique Transcription Evaluation for Named Entities*) proposée par (Ben Jannet *et al.*, 2015b) qui est fondée sur un modèle probabiliste estimant le risque qu'une erreur de RAP induise une erreur de REN. Elle s'appuie sur une comparaison de probabilités de présence d'éléments d'intérêts (des entités dans le cas de la REN, des concepts si on l'utilise dans le cas de la compréhension par exemple) dans les transcriptions de référence et dans les hypothèses des systèmes de RAP. ATENE est la moyenne de deux mesures élémentaires $ATENE_{DS}$ et $ATENE_I$. Elle est donnée par l'equation 1.

$$ATENE = -100\frac{ATENE_{DS} + ATENE_I}{2} \tag{1}$$

$ATENE_{DS}$, donnée par l'équation 2, est une mesure du risque d'erreurs de suppression et de substitution d'entités engendré par les erreurs de RAP et $ATENE_I$, donnée par l'équation 3, est une mesure du risque d'erreurs d'insertion d'entités engendré par les erreurs de RAP.

$$ATENE_{DS} = \frac{\sum_{i=1}^{N} \Delta_p(\text{début}_i) + \Delta_p(\text{fin}_i)}{2N} \tag{2}$$

Δ_p est la différence de probabilités calculée sur des mots se trouvant en début et fin d'entités et N le nombre d'entités ou de concepts.

$$ATENE_I = \frac{\sum_{i=1}^{N_S} \Delta_{PS}(S_i)}{N_S} \tag{3}$$

Δ_{PS} est la différence de risque d'insertion calculée sur des segments de parole ne contenant pas d'entités nommées et N_S le nombre de segments entre entités ou concepts.

ATENE a obtenu de meilleures corrélations que le WER, le NE-WER ou encore les mesures de pertes d'information (WIL, et P, R, F) entre les performances obtenues par les systèmes de RAP et celles des systèmes de REN sur des données des campagnes QUAERO et ETAPE (Ben Jannet, 2015).

Comme nous l'avons dit, la tâche de REN semble relativement proche d'une tâche de compréhension de la parole, tout au moins telle que cette dernière est le plus souvent envisagée, c'est à dire comme une tâche de repérage de mentions de concepts dans un texte (voir la sous-section 3.1). Nous nous attendons donc à ce que ATENE permette d'obtenir une bonne corrélation avec l'évaluation globale de la tâche de compréhension. Ceci constitue la première partie des expériences que nous présentons ici. La deuxième partie, consiste à vérifier à quel point ATENE est dépendante d'un modèle correspondant strictement à la tâche. Pour cela nous tentons d'établir une correspondance entre la tâche de REN et celle de compréhension.

3 Tâche et Données

3.1 Compréhension de la parole

La tâche de compréhension de la parole consiste à associer un *sens* à un signal de parole. Donner un sens signifie que l'information, l'intention qu'a exprimée le locuteur doit pouvoir être traitée par l'ordinateur. Il s'agit donc de transformer le signal de parole en une interprétation sémantique, un langage formel qui traduit pour l'ordinateur le sens porté par les paroles du locuteur.

Généralement cette tâche s'exécute en deux temps. Tout d'abord, le signal de parole est transcrit automatiquement en une chaîne lexicale. Puis des systèmes de compréhension traitent cette chaîne lexicale afin d'en extraire une interprétation sémantique.

Le choix de la représentation sémantique est essentiel et doit être adapté aux données à analyser. On peut remarquer que dans la littérature, chaque tâche de compréhension adopte une représentation qui lui est propre. Cette diversité s'explique par la diversité des données traitées dans chaque application et par le fait qu'il n'existe pas de représentation sémantique générique qui puisse répondre aux besoins de toutes les applications. Par conséquent, dans la pratique, chaque tâche de compréhension adopte une représentation sémantique *ad hoc*.

Dans cet article, nous nous intéressons plus particulièrement aux applications de type dialogue homme-machine. Dans ce cadre, la tâche de compréhension consiste souvent à rechercher des renseignements qui sont liés à une base de données (tâche de *slot-filling*). Dans ce cadre particulier, chacun s'accorde notamment sur le fait que les éléments de base de la représentation sémantique sont des *concepts* et que chaque concept est associé à une *valeur* (e.g. : corpus ATIS (Hemphill *et al.*, 1990), corpus MEDIA (Bonneau-Maynard *et al.*, 2009), corpus PORTMEDIA (Lefèvre *et al.*, 2012)).

3.2 MEDIA : réservation d'hôtels et informations touristiques

Nous avons choisi de travailler sur le corpus MEDIA qui a été créé dans le but explicite de fournir à la communauté un corpus annoté sémantiquement afin d'évaluer et de comparer les différents systèmes de compréhension de la parole.

Le corpus comprend 1 257 dialogues entre un système simulé par un humain et un utilisateur souhaitant se renseigner ou réserver un hôtel. Le corpus a été transcrit et annoté manuellement. Les annotations sémantiques se composent de *modes*, *concepts*, *valeurs* et *spécifieurs*. Nous avons choisi dans un premier temps de ne tenir compte que du niveau *concept*, qui dans MEDIA correspond à 74 étiquettes définies selon une ontologie du domaine. Ils peuvent par exemple être des concepts généraux (réponse, nombre, temps-date, ...) ou se référer directement à des objets de la potentielle base de données (chambre, hotel). Chaque tour de parole est alors subdivisé en séquences de mots qui sont : soit associées à un concept (la séquence de mot est alors appelée support du concept) ; soit associées à l'étiquette *null* si la séquence de mots ne porte pas de sens vis à vis de l'application. Un exemple de texte annoté est donné dans la figure 1. Les 17 693 tours de paroles utilisateurs sont répartis selon trois sous-corpus : l'apprentissage contient 12 916 énoncés, le développement (DEV_MEDIA) en contient 1 259 et finalement le test (TEST_MEDIA) est composé de 3 518 énoncés. Un système de reconnaissance automatique de la parole identique à celui présenté dans (Bougares *et al.*, 2013) a été appliqué sur le corpus MEDIA afin d'obtenir des transcriptions automatiques. Il s'agit d'un système

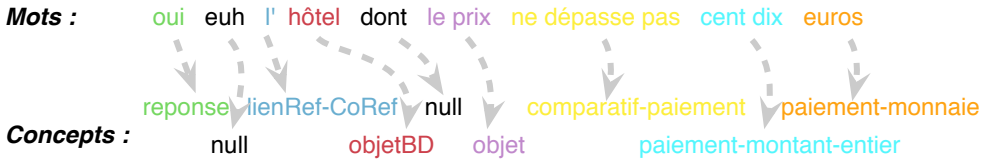

FIGURE 1 – Exemple d'interprétation sémantique en concept dans MEDIA

.

multi-passes fondé sur le projet CMU Sphinx [1], utilisant un modèle de langage quadrigramme adapté à MEDIA. Afin de générer différentes valeurs de WER, une segmentation manuelle ou automatique est utilisée et les paramètres du système ASR (poids du modèle de langage, poids d'insertion d'un mot, ...) ont été modulées. En conséquence, nous obtenons des transcriptions différentes qui présentent sur le DEV (4 systèmes) un WER oscillant entre 26,22% et 28,06% et sur le TEST (6 systèmes) entre 25,28% et 27,45%.

3.3 Implémentation du système de compréhension sur MEDIA

Dans (Hahn *et al.*, 2011), plusieurs systèmes de compréhension fondés sur différents algorithmes ont été implémentés et évalués. Les conclusions ont montré que le système fondé sur les champs conditionnels aléatoires (Conditional Random Field - CRF) (Lafferty *et al.*, 2001) était le plus performant. Récemment, plusieurs investigations de systèmes de compréhension de la parole fondés sur des réseaux de neurones ont été implémentés (*e.g.* (Mesnil *et al.*, 2015), (Liu & Lane, 2015), (Shi *et al.*, 2015)). Il a néanmoins été démontré dans (Vukotic *et al.*, 2015) que sur un corpus aussi complexe que MEDIA, le système fondé sur les CRF obtenait toujours les meilleurs résultats.

Nous avons donc choisi de ré-implémenter ce système en utilisant comme paramètres d'entrée : le mot, son étiquette sémantique, les 1 à 4 premières lettres du mot, les 1 à 4 dernières lettres du mot et également un paramètre indiquant l'absence ou la présence d'une majuscule sur la première lettre du mot. Pour obtenir différentes performances, nous utilisons plusieurs fichiers de paramètres afin que le CRF ne prenne pas en compte la totalité des paramètres d'entrée pour tous les systèmes mais seulement les 2 premiers, les 3 premiers... puis prenne en compte le contexte du mot plus ou moins élargi (0 à 2 mots avant, 0 à 2 mots après). Il en est de même pour le contexte de l'étiquette sémantique du mot. Cette dernière est issue d'un lexique obtenu manuellement. En effet dans MEDIA, le but du dialogue pour l'utilisateur est d'obtenir des informations qui sont stockées dans une base de données. Par conséquent, les noms de rues, de villes ou d'hôtels, les listes d'équipement de chambre, les types de nourriture, *etc.* sont connus. De plus, des mots plus généraux représentant les nombres, les jours, les mois sont également connus. Tous ces mots (spécifiques à la tâche SLU ou généraux) ont été rassemblés dans un lexique sémantique qui permet d'associer un mot à une étiquette sémantique.

1. http ://cmusphinx.sourceforge.net

4 Expériences

4.1 Méthodologie

Nous souhaitons dans un premier temps vérifier que nous pouvons utiliser la méthodologie proposée avec ATENE pour l'évaluation de système de RAP en vue d'une application de compréhension de la parole et obtenir de bonnes corrélations avec le CER [2], métrique utilisée ici pour évaluer la tâche de compréhension. Pour cela nous avons mis en place des expériences en suivant la méthodologie présentée dans (Ben Jannet, 2015). La première chose à faire est de développer les modèles nécessaires à l'utilisation de ATENE pour calculer les probabilités qui permettent d'estimer le risque d'erreurs. Un premier modèle permet d'estimer pour chaque mot sa probabilité d'être un début de support de concept-X ou non ; un deuxième permet d'estimer la probabilité pour un mot d'être une fin de support de concept-X ou non ; un dernier modèle permet d'estimer la probabilité qu'un mot soit n'importe où dans un support de concept ou non. Les deux premier modèles permettent de calculer $ATENE_{DS}$ en calculant le risque d'erreur induit par la RAP. Ce risque d'erreur est estimé par la différence entre les marges [3] calculées sur la transcription et la transcription automatique, comme indiqué dans l'équation 2. $ATENE_i$ est calculé en s'appuyant sur le troisième modèle qui estime la probabilité pour chaque mot d'être ou non dans un support de concept. Ce modèle n'est appliqué que sur les segments de parole hors concepts.

Nous avons appris différents modèles sur les données d'apprentissage du corpus MEDIA tel que présenté dans la section 3 : en utilisant l'ensemble des concepts détaillés ce qui représente 74 concepts y compris le concept NULL, en utilisant uniquement les têtes des concepts (par exemple COMMAND pour COMMAND-TACHE et COMMAND-DIAL) ce qui représente 22 concepts. Ceci représente six modèles qui s'appuient tous sur les mêmes traits : les mots dans une fenêtre $[-2, +2]$ et les préfixes et suffixes jusqu'à quatre caractères du mot courant.

Pour répondre à notre deuxième question sur la possibilité ou non d'utiliser des modèles appris sur une tâche différente, nous avons utilisé des modèles appris sur les données QUAERO [4] annotées en entités nommées. Nous n'avons considéré dans ces modèles que les entités pouvant potentiellement se rapprocher de certaines têtes de concept du modèle de compréhension : les dates (qui correspondent au concept TEMPS), les lieux désignant des villes (qui correspondent au concept LOCALISATION), les montants (qui correspondent aux concepts NOMBRE, SEJOUR, PAIEMENT et NOMBRENONDIGIT) , les organisations (qui correspondent aux concepts HOTEL et NOM) et les personnes.

Ici nous nous intéressons à la capacité des métriques d'évaluation des différents systèmes de RAP en fonction de la qualité de leur sortie étant donné la tâche de compréhension. Autrement dit, nous ne cherchons pas à dire tel système commet moins d'erreur que tel autre mais plutôt tel système comment moins d'erreurs ayant impact sur la tâche SLU que tel autre. Pour cela, nous comparons pour chaque métrique les classements des systèmes de RAP obtenus selon les performances de compréhension (donc selon le résultat obtenu en terme de CER) et les rangs des sytèmes de RAP selon les scores fournis par les métriques d'évaluation de RAP. Nous utilisons pour cela la corrélation de Kendall qui

2. Le Concept Error Rate, ou CER est une métrique d'évaluation largement utilisée en compréhension de la parole et qui s'apparente au WER. Dans le calcul du CER, on compare la liste des concepts de référence avec la liste des concepts fournis par le système de compréhension automatique. Il est à noter que le nombre de concepts dans une phrase n'est pas obligatoirement égal au nombre de mots.

3. La *Marge* est l'écart de probabilités entre le label attendu et le meilleur label pour un modèle donné (Ben Jannet, 2015).

4. Nous utilisons ces données car elles sont disponibles gratuitement pour la recherche académique, voir `http://catalog.elra.info/product_info.php?products_id=1195`.

reflète le degré de concordance et de discordance entre les rangs de deux classements. Cette mesure donne des valeurs comprises entre -1 et +1 dont la valeur absolue indique la puissance de corrélation entre les deux variables testées. Il est à noter que les corrélations ne sont pas calculables quand une des mesures comparées ne fournit que des scores identiques. Hors cela arrive parfois, en particulier avec CER et WER-NE. Nous nous sommes donc restreints aux dialogues où cette situation ne se produit pas.

Métrique	Dev.				Test.			
	Kendall	IC. 95%	#ASR	#Tests	Kendall	IC. 95%	#ASR	#Tests
WER	0,853	[0,598 ;1]	4	294	0,837	[0,603 ;1]	6	939
1-F	0,869	[0,641 ;1]	4	294	0,851	[0,656 ;1]	6	939
WER-NE	0,771	[0,425 ;1]	4	294	0,841	[0,545 ;1]	6	939
ATENE-SLU	0,885	[0,701 ;1]	4	294	0,858	[0,661 ;1]	6	939
ATENE-TOP	0,866	[0,612 ;1]	4	294	0,857	[0,661 ;1]	6	939
ATENE-NE	0,852	[0,563 ;1]	4	294	0,840	[0,613 ;1]	6	939

FIGURE 2 – Corrélation de Kendall moyenne avec intervalle de confiance à 95% entre les différentes métriques et le CER. #ASR indique le nombre de systèmes ASR comparés, #Tests indique le nombre de dialogues sur lesquels la corrélation de rang est calculée. ATENE-SLU réfère à la métrique ATENE utilisant les modèles appris sur les données d'apprentissage MEDIA de la tâche SLU, ATENE-TOP est identique mais utilise uniquement les têtes de concept et ATENE-NE renvoie à l'utilisation des modèles appris sur le corpus QUAERO pour la tâche NER.

4.2 Résultats

Le tableau 2 montre les différents résultats obtenus. Il contient les coefficients de corrélation (τ de Kendall) associés aux intervalles de confiance à 95 %. Les résultats sont donnés pour le corpus de développement et le corpus de test. Nous indiquons ce qui a été obtenu pour chacune des métriques évaluées : le WER, ATENE-SLU, c'est à dire ATENE construit avec les données adaptées à la tâche SLU, les données MEDIA, ATENE-TOP qui s'appuie sur les têtes de concept, ATENE-NE, construit avec les données du corpus QUAERO mais uniquement certaines entités, la F-mesure ou plutôt son complémentaire (1-F) et le NE-WER.

On constate que ATENE-SLU a effectivement une légèrement meilleure corrélation que WER. La différence est loin d'être aussi importante que ce qu'elle était sur les tâches d'EN telles que rapportées dans (Ben Jannet, 2015) mais reste présente. Il est intéressant de noter que l'intervalle de confiance est mieux resserré avec ATENE-SLU qu'avec le WER, indiquant que ATENE-SLU est moins bruitée, et donc probablement plus utile dans le cadre d'un développement de système de RAP avec une faible quantité de données de développement disponibles. On peut voir que la F-mesure est de façon inattendu un meilleur estimateur de la qualité des systèmes de RAP que le WER mais n'atteint pas le niveau d'ATENE. L'approche WER-NE ne semble pas produire de résultats particulièrement intéressants. La non-prise en compte des effets induits par les insertions de mot par les systèmes de RAP la rend tout aussi peu fiable que dans le cadre de la REN. La légère meilleure performance de ATENE-SLU par rapport à ATENE-TOP semble indiquer qu'ATENE a besoin de modèles relativement précis plutôt que larges. ATENE-NE confirme cette tendance. En effet, des modèles non spécifiques à la tâche ne permettent pas d'obtenir une corrélation raisonnable.

5 Conclusion et perspectives

Nous avons présenté une série d'expériences menées pour vérifier à quel point une métrique d'évaluation de systèmes de RAP peut corréler avec le résultat global d'une tâche de SLU. Nous avons comparé différentes métriques dont la métrique la plus utilisée, le taux d'erreurs mots (WER), et ATENE, une métrique récemment proposée pour l'évaluation de systèmes de RAP en vue d'une tâche de reconnaissance d'entités nommées. Cette dernière métrique a été implémentée d'une part en utilisant les données liées à la tâche considérée (les données MEDIA) mais aussi en utilisant les données liées à une tâche de reconnaissance d'entités nommées (les données QUAERO). Ce dernier cas avait pour objectif de mesurer la dépendance de ATENE aux données de la tâche. ATENE, dans sa forme la plus complète et la plus adaptée à la tâche, donne le meilleur résultat, c'est à dire un coefficient de corrélation supérieur à 0,85 avec un intervalle de confiance plus resserré qu'avec le WER. Nous allons poursuivre nos travaux dans deux directions. La première utilisation d'ATENE peut être de l'utiliser pour le *tuning* d'un système de RAP. Toutefois, nous considérons qu'améliorer ATENE, ou n'importe quelle autre métrique, pour étudier la corrélation entre les performances de RAP et celles de SLU passe par la définition d'une meilleure métrique SLU qui prenne en compte la totalité du problème, non seulement la détection des concepts mais aussi de leur mode et leur valeur.

Remerciements

Ce travail a été financé partiellement par le projet VERA - ANR 12 BS02 006 04.

Références

BEN JANNET M. A. (2015). *Évaluation adaptative des systèmes de transcription en contextes applicatifs*. PhD thesis, Université Paris Sud.

BEN JANNET M. A., GALIBERT O., ADDA-DECKER M. & ROSSET S. (2015a). How to evaluate asr output for named entity recognition ? In *16th Annual Conference of the International Speech Communication Association (Interspeech'15)*.

BEN JANNET M. A., GALIBERT O., ADDA-DECKER M. & ROSSET S. (2015b). How to evaluate asr output for named entity recognition ? In *Interspeech*, Dresden, Germany.

BONNEAU-MAYNARD H., QUIGNARD M. & DENIS A. (2009). Media : a semantically annotated corpus of task oriented dialogs in french. *Language Resources and Evaluation*, **43**(4), 329–354.

BOUGARES F., DELÉGLISE P., ESTEVE Y. & ROUVIER M. (2013). Lium asr system for etape french evaluation campaign : experiments on system combination using open-source recognizers. In *Text, Speech, and Dialogue*, p. 319–326 : Springer.

FAVRE B., CHEUNG K., KAZEMIAN S., LEE A., LIU Y., MUNTEANU C., NENKOVA A., OCHEI D., PENN G., TRATZ S., VOSS C. & ZELLER F. (2013). Automatic Human Utility Evaluation of ASR Systems : Does WER Really Predict Performance ? In *Interspeech, Lyon (France)*.

GAROFOLO J. S., VOORHEES E. M., AUZANNE C. G., STANFORD V. M. & LUND B. A. (1999). 1998 trec-7 spoken document retrieval track overview and results. In *Broadcast News Workshop'99 Proceedings*, p. 215 : Morgan Kaufmann Pub.

HAHN S., DINARELLI M., RAYMOND C., LEFEVRE F., LEHNEN P., DE MORI R., MOSCHITTI A., NEY H. & RICCARDI G. (2011). Comparing stochastic approaches to spoken language understanding in multiple languages. *Audio, Speech, and Language Processing, IEEE Transactions on*, **19**(6), 1569–1583.

HEMPHILL C. T., GODFREY J. J. & DODDINGTON G. R. (1990). The atis spoken language systems pilot corpus. In *Proceedings of the DARPA speech and natural language workshop*, p. 96–101.

LAFFERTY J. D., MCCALLUM J. D. & PEREIRA F. C. N. (2001). Conditional random fields : probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, San Francisco, CA, USA.

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 *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1 : JEP*, p. 779–786 : ATALA/AFCP.

LIU B. & LANE I. (2015). Recurrent neural network structured output prediction for spoken language understanding. In *Proc. NIPS Workshop on Machine Learning for Spoken Language Understanding and Interactions*.

MCCOWAN I. A., MOORE D., DINES J., GATICA-PEREZ D., FLYNN M., WELLNER P. & BOURLARD H. (2004). *On the use of information retrieval measures for speech recognition evaluation*. Rapport interne, IDIAP.

MESNIL G., DAUPHIN Y., YAO K., BENGIO Y., DENG L., HAKKANI-TUR D., HE X., HECK L., TUR G., YU D. *et al.* (2015). Using recurrent neural networks for slot filling in spoken language understanding. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, **23**(3), 530–539.

MILLER G. A. (1955). Note on the bias of information estimates. *Information theory in psychology : Problems and methods*, **2**, 95–100.

MORRIS A. C., MAIER V. & GREEN P. (2004). From wer and ril to mer and wil : improved evaluation measures for connected speech recognition. In *INTERSPEECH*.

MUNTEANU C., BAECKER R., PENN G., TOMS E. & JAMES D. (2006). The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, p. 493–502 : ACM.

PALLETT D. S. (2003). A look at nist's benchmark asr tests : past, present, and future. In *ASRU'03*.

PRZYBOCKI M. A., FISCUS J. G., GAROFOLO J. S. & PALLETT D. S. (1999). 1998 hub-4 information extraction evaluation. In *Proc. DARPA Broadcast News Workshop,(Herndon, Va, USA)*, p. 13–18.

RICCARDI G. & GORIN A. L. (1998). Stochastic language models for speech recognition and understanding. In *ICSLP*.

SHI Y., YAO K., CHEN H., PAN Y.-C., HWANG M.-Y. & PENG B. (2015). Contextual spoken language understanding using recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, p. 5271–5275 : IEEE.

VUKOTIC V., RAYMOND C. & GRAVIER G. (2015). Is it time to switch to word embedding and recurrent neural networks for spoken language understanding ? In *InterSpeech*.

WANG Y.-Y., ACERO A. & CHELBA C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In *ASRU'03*, p. 577–582 : IEEE.