Comparaison de listes d'erreurs de transcription automatique de la parole : quelle complémentarité entre différentes métriques ?

Olivier Galibert[1] Juliette Kahn[1] Sophie Rosset[2]
(1) LNE, F-78190 Trappes, France
(2) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
olivier.galibert@lne.fr, juliette.kahn@lne.fr, rosset@limsi.fr

RÉSUMÉ

Le travail que nous présentons ici s'inscrit dans le domaine de l'évaluation des systèmes de reconnaissance automatique de la parole en vue de leur utilisation dans une tâche aval, ici la reconnaissance des entités nommées. Plus largement, la question que nous nous posons est "que peut apporter une métrique d'évaluation en dehors d'un score ?". Nous nous intéressons particulièrement aux erreurs des systèmes et à leur analyse et éventuellement à l'utilisation de ce que nous connaissons de ces erreurs. Nous étudions dans ce travail les listes ordonnées d'erreurs générées à partir de différentes métriques et analysons ce qui en ressort. Nous avons appliqué la même méthode sur les sorties de différents systèmes de reconnaissance de la parole. Nos expériences mettent en évidence que certaines métriques apportent une information plus pertinente étant donné une tâche et transverse à différents systèmes.

ABSTRACT

Comparing error lists for ASR systems : contribution of different metrics.

The work presented here is concerned by the evaluation of automatic recognition systems used as a first step in a broader task on named entity recognition in spoken data. More precisely, the question we are asking is "what can an evaluation metric bring besides a score ?. We are in particular interested in the errors done by the systems and their analysis, and possibly the use of what we can learn of these errors. In this work we produced ordered error lists based on various metrics and analysed them. Our experiments show that some metrics provide a more pertinent information for the NER task that is almost system-independent.

MOTS-CLÉS : Reconnaissance automatique de la parole, Métriques d'évaluation, Analyse d'erreurs.

KEYWORDS: Automatic speech recognition, Metrics, Error analysis.

1 Introduction

Le travail que nous présentons ici [1] s'inscrit dans le domaine de l'évaluation des systèmes de reconnaissance automatique de la parole (RAP). Nous souhaitons évaluer ces systèmes sachant qu'ils seront utilisés dans une tâche aval et adapter l'évaluation de la transcription à ce type d'usage. Au vu des données à notre disposition, nous avons choisi ici comme tâche aval la reconnaissance des entités nommées (REN). Plus largement, nous nous interrogeons sur l'apport d'une métrique d'évaluation au

1. Cet article est une adaptation et extension de l'abstract soumis à LREC 2016.

delà de l'optimisation du score. Nous nous intéressons particulièrement aux erreurs des systèmes et à leur analyse, et éventuellement à l'utilisation de ce que nous connaissons de ces erreurs.

Les erreurs produites par les systèmes de RAP peuvent avoir un impact plus ou moins important selon leur fréquence, mais aussi selon l'utilisation qui est faite de la sortie du système (voir par exemple (Comas & Turmo, 2009) pour des systèmes de réponse précise à des questions en langue naturelle ou encore (Dinarelli & Rosset, 2011) pour des systèmes de REN). Ainsi comprendre et éventuellement anticiper les types des erreurs ou au moins leur impact sur la tâche qui suit celle de la RAP est un élément important pour améliorer la robustesse de ces systèmes, tant de RAP que de REN.

Plusieurs travaux ont porté sur l'analyse des erreurs produites par les systèmes de RAP. Ils avaient comme objectifs soit d'améliorer les systèmes de RAP eux-mêmes (Boháč *et al.*, 2012; Dufour & Esteve, 2008) soit de détecter automatiquement les erreurs (Ghannay *et al.*, 2015). Il est à noter que tenir compte des résultats de ces travaux restait à la charge des systèmes appliqués en aval.

Une grande partie des travaux qui se sont intéressés à l'étude des erreurs résiduelles des systèmes de RAP l'ont principalement fait dans le cadre de comparaisons entre erreurs produites par des systèmes et erreurs produites par des humains (Scharenborg, 2007; Lippmann, 1997). Ces études mettent en évidence que si les systèmes sont devenus plutôt très performants, ils ne sont pas encore en mesure de prendre en compte toutes les variations acoustiques observées. Par ailleurs, ces résultats montrent que les humains lorsqu'ils transcrivent (ou écoutent) de la parole obtiennent des performances cinq à six fois supérieures à celles des systèmes (Vasilescu *et al.*, 2012). Ces travaux ont permis de faire émerger des taxonomies d'erreurs indiquant que certains mots sont plus sujets à erreur que d'autres, en particulier les mots courts, pauvres d'un point de vue acoustique, ou encore les homophones courts potentiellement ambigus comme le verbe *a* et la préposition *à* (Adda-Decker, 2006). D'autres travaux se sont penchés sur la classification des erreurs. Par exemple, (Goryainova *et al.*, 2014) ont étudié les erreurs étant donné des classes morpho-syntaxiques de mots et (Santiago *et al.*, 2015) ont proposé une typologie d'erreurs fondée sur des critères syntaxiques et prosodiques ou encore (Rena Nemoto & Adda-Decker, 2008)) comparent des classifications fondées sur des études perceptives et des classifications automatiques des erreurs fréquentes sur des homophones.

Notre hypothèse de travail est qu'une métrique d'évaluation peut permettre de donner des indications sur les erreurs produites par un système de RAP. Plus précisément, nous considérons qu'une métrique doit pouvoir fournir une information sur la gravité d'une erreur, au sens de l'importance de celle-ci étant donnée la tâche globale. Cette gravité peut être dépendante de différents critères comme la fréquence certes, mais aussi son impact sur la capacité des systèmes intervenant en aval de la reconnaissance de la parole à fournir le meilleur résultat possible.

La section 2 présente notre proposition en y associant une brève description de différentes métriques utilisées pour l'évaluation des systèmes de RAP en contexte de REN. La section 3 présente les expériences que nous avons réalisées, les résultats obtenus et discute ceux-ci. Enfin la dernière section présente un bilan de cette expérience ainsi que quelques perspectives ouvertes.

2 Proposition

Considérant qu'obtenir une liste des erreurs et de leur importance passe par l'utilisation d'une métrique d'évaluation, nous commençons dans cette section par décrire quelques unes des métriques existantes. Puis nous détaillons notre proposition.

2.1 Évaluation des systèmes de reconnaissance automatique de la parole

La métrique la plus utilisée est le taux d'erreur mots (WER pour *Word Error Rate* (Pallett, 2003). Cette métrique compte les erreurs et normalise ce décompte par la taille (en nombre de mots) de la référence. Il existe différents types d'erreur, la substitution, la suppression et l'insertion, qui sont déterminés par un alignement de Levenshtein (Levenshtein, 1966) entre la référence (transcription manuelle) et l'hypothèse (transcription automatique). Le WER est donc une métrique fondée sur une énumération simple d'erreurs qui au final considère toutes les erreurs comme étant d'égale importance. Avec cette métrique plus une erreur est fréquente plus elle est grave.

Lorsque la RAP est la première étape d'une tâche plus globale, certaines études ont montré que le WER n'était pas toujours la métrique offrant la meilleure corrélation entre les performances du système de RAP et les performances obtenues sur la tâche globale comme dans le cas de la REN et de la recherche d'information (Garofolo *et al.*, 2000), de la traduction automatique (He *et al.*, 2011) ou encore de la compréhension de la parole (Wang *et al.*, 2003).

Plusieurs alternatives au WER existent. (Miller, 1955) a développé une mesure de la perte d'information occasionnée par les erreurs de RAP. Cette mesure, appelée le RIL (*Relative Information Loss*), est fondée 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. Elle est représentée en termes d'entropie de Shannon. Par la suite, (Morris *et al.*, 2004) ont introduit le WIL (*Word Information Loss*) qui est une approximation du RIL. (Morris *et al.*, 2004) et (McCowan *et al.*, 2004) ont montré que ces deux métriques, RIL et WIL, présentent un intérêt lorsque le taux d'erreur est supérieur à 50%. 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 générale 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. Comme nous nous intéressons à la reconnaissance d'entités nommées sur des données de parole, nous retenons la proposition de (Garofolo *et al.*, 1999) qui ont proposé de calculer un WER mais limité aux mots de la référence qui sont présents dans une entité nommée (NE-WER, *Named Entity Word Error Rate*). Un inconvénient de cette métrique est qu'elle ignore les mots insérés ou substitués en dehors des entités nommées mais qui peuvent conduire à une fausse alarme (détection erronée d'une entité nommée). Plus récemment et toujours pour évaluer la RAP dans le contexte de la REN, la métrique ATENE a été proposée par (Ben Jannet *et al.*, 2015). Cette métrique est fondée sur un modèle probabiliste qui estime le risque qu'une erreur de RAP induise une erreur de REN. 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 obtenus par les systèmes de RAP et celles des systèmes de REN.

Si disposer d'une métrique permettant d'estimer la qualité d'un système de RAP en tâche aval est intéressant, cela ne permet toutefois pas d'obtenir une liste des erreurs les plus coûteuses ou les plus fréquentes. Une telle liste est cependant utile pour améliorer les systèmes de RAP (Dufour & Esteve, 2008). Notre objectif est donc de pouvoir produire une liste des erreurs avec leur fréquence mais aussi, et surtout, une classification de ces erreurs en fonction de leur coût étant donnée une tâche. Les mesures générales comme le RIL et le WIL ne permettent pas de quantifier l'impact d'erreurs spécifiques, elles donnent un point de vue global sur la qualité d'un système de RAP. Nous avons décidé de nous appuyer sur ATENE qui a montré une corrélation plus importante que le WER ou le triplet P, R et F pour ce travail (Ben Jannet *et al.*, 2015).

2.2 Listes d'erreurs

Pour générer une liste d'erreurs qui quantifie l'impact des erreurs selon la tâche de REN, la première étape est de générer une liste d'erreurs. La méthodologie utilisée pour calculer un WER, en produisant un alignement entre les mots de la références et ceux de l'hypothèse, produit une telle liste. Il s'agit d'ailleurs de la seule métrique permettant cela. Cet alignement classifie les erreurs en insertions, suppressions et substitutions. Dénombrer les erreurs de cette liste permet d'obtenir une liste d'erreurs ordonnées selon leur importance telle que considérée par la métrique WER.

Dans ce travail, nous avons décidé de conserver cette liste d'erreurs mais de calculer pour chaque erreur son poids en fonction de son impact sur la métrique ATENE. Pour chaque occurrence d'une erreur, la transformation correspondant à l'erreur réalisée dans la référence est appliquée pour créer une nouvelle hypothèse. Le score de cette hypothèse est calculé avec la métrique ATENE. Cette métrique se comportant comme un taux d'erreur, les scores de chaque occurence sont cumulés pour obtenir le poids final qui lui sera associé. Le score peut en pratique être obtenu par un petit nombre de calculs autour du site de chaque instance d'erreur.

Le résultat de ces calculs permet alors de trier la liste et d'ordonner les erreurs selon leur importance. L'outil d'évaluation *sclite* qui implémente le WER fournit une liste d'erreurs ordonnée selon leur fréquence dans le fichier *dtl* (liste WER). Nous avons créé une liste équivalente en utilisant le résultat de l'impact de l'erreur sur ATENE comme élément d'importance (liste Atene). Afin d'affiner nos comparaisons, nous avons également produit deux autres listes : la première, inspirée par le NE-WER, ne considère que les occurrences d'erreurs apparaissant dans une entité nommée (liste In) et une seconde ne prenant en compte que les erreurs dans ou au contact d'une entité nommée (liste Near).

3 Expérimentations et résultats

3.1 Données

Pour nos expérimentations, nous avons utilisé les données de tests de la campagne ETAPE (Galibert *et al.*, 2014). Ces données, présentées dans le tableau 1, consistent en 15 émissions radiophoniques différentes, transcrites manuellement et par cinq systèmes de RAP (les participants à la campagne) ainsi que par un système rover.

	Test
Mots	115 803
Entités	5 933

	RAP-1	RAP-2	RAP-3	RAP-4	RAP-5	ROVER
WER	22,3	25,7	26,6	30,4	36,7	28,68

TABLE 1 – Description du corpus ETAPE : quantité de données et performance des systèmes

3.2 Méthodologie

Nous avons généré les listes des erreurs les plus impactantes selon chacune des métriques (ATENE, WER, In et Near) pour chaque système. Ces listes ont été générées avec deux seuils différents : la

Liste-10 conserve pour chaque système et chaque métrique les 10 erreurs les plus graves et inscrit pour chacune de ces erreurs le rang obtenu pour chaque métrique, la Liste-100 fait de même avec les 100 erreurs les plus graves. Si aucun recouvrement d'erreur n'existait entre les systèmes et les métriques, c'est à dire si chaque système et chaque métrique donnait des erreurs différentes, alors la fusion des Liste-10 devrait contenir $10\,(erreurs) \times 6\,(systèmes) \times 4\,(métriques) = 240$ entrées. Pourtant, avec les recouvrements, la fusion des Liste-10 se composent de 47 entrées. De la même manière, pour Liste-100, le nombre maximal d'entrées est de $100 \times 6 \times 4 = 2400$ entrées. Nous en observons en réalité 733. Ceci montre déjà que le recouvrement est important.

Nous souhaitons comparer ces listes afin de vérifier les hypothèses suivantes :

1. les listes générées par les quatre métriques pour un même système sont différentes. Si cette hypothèse est vérifiée, cela signifie que l'impact mesuré des erreurs d'un même système n'est pas le même selon la métrique utilisée ;
2. l'ordre d'importance des erreurs proposé par une métrique est équivalent quel que soit le système. Si cette hypothèse est vérifiée, cela signifie que la métrique est cohérente.

Pour vérifier ces deux hypothèses nous utilisons des mesures de corrélations de rang de Spearman (Spearman, 1904). Cette mesure donne des valeurs comprises entre -1 et +1 indiquant la puissance de corrélation entre les deux variable testées. Si la valeur est élevée ($\geq 0,8$), cela signifie que l'ordre des erreurs est le même. Au contraire, si la valeur absolue de la corrélation de rang est basse, cela signifie que les listes sont différentes et que ce ne sont pas les mêmes erreurs qui sont considérées comme importantes. Une valeur très négative indique que les listes sont en ordre inverse.

Une autre façon d'estimer la qualité de ces listes est de comparer leur contenu et les rangs des erreurs relevés, notamment en rapport avec la tâche. Si cette analyse ne peut être que partielle, elle peut donner des indications intéressantes.

3.3 Comparaisons de liste

Les corrélations de rang ont été calculées deux à deux pour chaque système et présentées sous forme de matrices sur les figures 1 et 2 respectivement pour Liste-10 et Liste-100. Les matrices se lisent de la manière suivante : plus une corrélation est forte ($R > 0,8$), plus la case est verte. Plus une corrélation est faible ($R < 0,3$), plus la case est rouge.

Analysons dans un premier temps la cohérence des métriques en fonction des systèmes afin de vérifier si pour une métrique donnée, les listes générées sont équivalentes pour tous les systèmes. La moyenne de corrélation de WER est respectivement de 0,96 (σ = 0,04) et de 0,91 (σ = 0,06) pour Liste-10 et Liste-100. Pour ATENE, la moyenne de corrélation est de 0,85 (σ = 0,10) et de 0,83 (σ = 0,09) Liste-10 et Liste-100. Pour Near, si pour Liste-10, la corrélation moyenne est de 0,84 (σ = 0,14), elle descend à 0,72 (σ = 0,16) sur Liste-100. Enfin, pour In, la moyenne des corrélations est de 0,66 (σ = 0,22) et de 0,64 (σ = 0,23) pour Liste-10 et Liste-100. Nous observons donc que les listes WER et Atene sont les plus cohérentes entre systèmes avec une corrélation moyenne supérieure à 0,8. En revanche la métrique In est celle qui semble la moins cohérentes avec une corrélation inférieure à 0,7. Les métriques WER et Atene quantifient donc les erreurs de manière très ressemblante d'un système à l'autre, et une prise en compte de la liste d'erreur d'un système RAP donné dans le système REN devrait être robuste au changement de système RAP.

Notre seconde question est de savoir si les listes d'erreurs sont différentes selon les métriques afin d'estimer leur complémentarité. Concernant le WER, pour Liste-10, sa corrélation moyenne avec

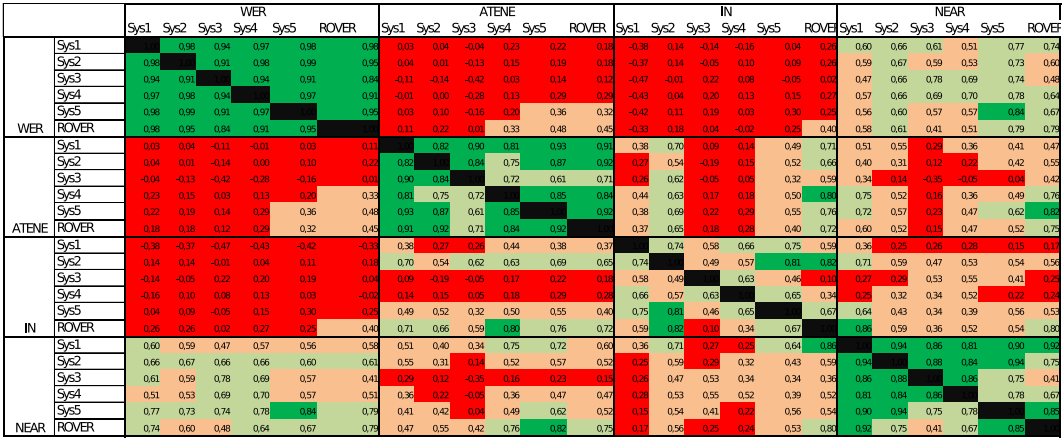

FIGURE 1 – Matrice de corrélation des mesures WER, ATENE, IN et Near pour les 5 systèmes de RAP et le ROVER sur Liste-10

ATENE est de 0,10 (σ = 0,19), de 0,03 (σ = 0,23) avec In et de 0,64 (σ = 0,10) avec Near. La même tendance s'observe avec Liste-100 : sa corrélation moyenne est de -0,17 (σ = 0,10) avec ATENE, de 0,14 (σ = 0,15) avec In et de 0,57 (σ = 0,09) avec Near. La corrélation très faible entre WER et In indique que les mots à l'intérieur des entités sont vraiment spécifiques par rapport à la langue globale. En revanche, WER corrèle mieux avec une métrique qui met en avant les erreurs autour des EN comme Near. Ceci est vraisemblablement dû au fait d'ajouter les mots autour des EN dilue leur spécificité en se rapprochant de la liste globale de WER.

Si on observe les corrélations entre ATENE et d'autres métriques, on constate une très faible corrélation entre WER et ATENE (0,10) montrant que ces deux mesures mettent en avant des erreurs différentes. Sa corrélation moyenne est un peu plus forte avec Near (0,40, σ = 0,11) et In (0,41, σ = 0,25) pour Liste-10. ATENE semble donc bien fournir des informations très différentes du WER grâce à sa prise en compte de la REN, le rapprochant de In mais allant plus loin.

3.4 Analyse qualitative

Voyons à présent plus en détail les erreurs jugées comme impactantes par ces différentes métriques.

Le tableau 2 contient quelques exemples extraits des différentes listes. Pour chacun des types d'erreur (suppressions, insertions et substitutions) il inclut l'erreur la plus importante pour chaque liste [2].

Comme nous pouvons le voir, la suppression considérée comme étant la plus importante par ATENE concerne la préposition *à*. Cette préposition est en général un bon indice pour la REN et nous pouvons voir que dans la référence ETAPE, 68% des occurrences de *à* se retrouvent devant une entité nommée. Il est probable que la suppression d'un tel mot a une conséquence sur un système de REN. La suppression de ce mot est également relativement importante pour le WER, ne serait-ce qu'à cause de sa fréquence (209 instances, 1,1% des erreurs). En revanche, nous pouvons constater que cette

2. Les listes d'erreurs ordonnent toutes les erreurs quel que soit leur type. Ainsi la première erreur de type X peut se retrouver au-delà de la première position dans la liste générale.

		WER						ATENE						IN						NEAR					
		Sys1	Sys2	Sys3	Sys4	Sys5	ROVER	Sys1	Sys2	Sys3	Sys4	Sys5	ROVER	Sys1	Sys2	Sys3	Sys4	Sys5	ROVER	Sys1	Sys2	Sys3	Sys4	Sys5	ROVER
	Sys1		0,96	0,94	0,95	0,96	0,88	-0,37	-0,08	-0,25	-0,08	-0,03	0,02	-0,02	0,09	0,18	0,01	0,23	0,27	0,63	0,63	0,62	0,55	0,66	0,52
	Sys2	0,96		0,93	0,93	0,92	0,86	-0,15	-0,38	-0,25	-0,19	-0,16	-0,06	-0,08	0,20	0,25	0,02	0,21	0,23	0,53	0,73	0,66	0,59	0,68	0,52
	Sys3	0,94	0,93		0,90	0,87	0,76	-0,30	-0,25	-0,34	-0,25	-0,17	-0,12	-0,13	-0,02	0,25	0,08	0,12	0,18	0,54	0,61	0,73	0,62	0,58	0,43
	Sys4	0,95	0,93	0,90		0,87	0,77	-0,17	-0,25	-0,30	-0,37	-0,16	-0,05	-0,13	-0,10	0,13	0,23	0,08	0,14	0,50	0,54	0,60	0,71	0,54	0,43
	Sys5	0,96	0,92	0,87	0,87		0,92	-0,09	-0,22	-0,19	-0,15	-0,19	-0,08	-0,06	0,01	0,19	0,10	0,40	0,29	0,49	0,54	0,56	0,53	0,71	0,53
WER	ROVER	0,88	0,86	0,76	0,77	0,92		-0,03	-0,12	-0,11	-0,03	-0,03	0,03	0,11	0,10	0,28	0,19	0,40	0,47	0,46	0,48	0,45	0,44	0,67	0,67
	Sys1	-0,37	-0,15	-0,30	-0,17	-0,09	-0,03		0,88	0,82	0,82	0,86	0,84	0,30	0,31	0,12	0,26	0,35	0,39	0,02	0,08	-0,10	0,05	0,17	0,30
	Sys2	-0,08	-0,38	-0,25	-0,25	-0,22	-0,12	0,88		0,74	0,77	0,78	0,80	0,32	0,14	0,06	0,13	0,22	0,29	0,22	-0,10	-0,09	-0,01	0,05	0,18
	Sys3	-0,25	-0,25	-0,34	-0,30	-0,19	-0,11	0,82	0,74		0,78	0,73	0,73	0,39	0,32	0,10	0,15	0,30	0,40	0,10	0,04	-0,18	-0,10	0,04	0,13
	Sys4	-0,08	-0,19	-0,25	-0,37	-0,15	-0,03	0,82	0,77	0,78		0,73	0,74	0,46	0,34	0,18	0,18	0,35	0,33	0,21	0,05	-0,04	-0,08	0,12	0,17
	Sys5	-0,03	-0,16	-0,17	-0,16	-0,19	-0,03	0,86	0,78	0,73	0,73		0,88	0,42	0,28	0,22	0,24	0,36	0,37	0,21	0,06	-0,02	0,00	0,12	0,21
ATENE	ROVER	0,02	-0,06	-0,12	-0,05	-0,08	-0,03	0,84	0,80	0,73	0,74	0,88		0,44	0,31	0,18	0,28	0,38	0,41	0,24	0,11	0,01	0,06	0,21	0,25
	Sys1	-0,02	-0,08	-0,13	-0,13	-0,06	0,11	0,30	0,32	0,39	0,46	0,42	0,44		0,61	0,56	0,69	0,70	0,73	0,39	0,17	0,08	0,16	0,29	0,40
	Sys2	0,09	0,20	-0,02	-0,10	0,01	0,10	0,31	0,14	0,32	0,34	0,28	0,31	0,61		0,53	0,49	0,54	0,51	0,41	0,44	0,16	0,12	0,27	0,39
	Sys3	0,18	0,25	0,25	0,13	0,19	0,28	0,12	0,06	0,10	0,18	0,22	0,18	0,56	0,53		0,47	0,51	0,57	0,47	0,39	0,46	0,31	0,35	0,38
	Sys4	0,01	0,02	0,08	0,23	0,10	0,19	0,26	0,13	0,15	0,18	0,24	0,28	0,69	0,49	0,47		0,39	0,44	0,38	0,22	0,25	0,53	0,29	0,37
	Sys5	0,23	0,21	0,12	0,08	0,40	0,40	0,35	0,22	0,30	0,35	0,36	0,38	0,70	0,54	0,51	0,39		0,77	0,46	0,31	0,27	0,18	0,63	0,58
IN	ROVER	0,27	0,23	0,18	0,14	0,29	0,47	0,39	0,29	0,40	0,33	0,37	0,41	0,73	0,51	0,57	0,44	0,77		0,49	0,36	0,32	0,29	0,50	0,70
	Sys1	0,63	0,53	0,54	0,50	0,49	0,46	0,02	0,22	0,10	0,21	0,21	0,24	0,39	0,41	0,47	0,38	0,46	0,49		0,78	0,76	0,73	0,75	0,60
	Sys2	0,63	0,73	0,61	0,54	0,54	0,48	0,08	-0,10	0,04	0,05	0,06	0,11	0,17	0,44	0,39	0,22	0,31	0,36	0,78		0,77	0,70	0,72	0,56
	Sys3	0,62	0,66	0,73	0,60	0,56	0,45	-0,10	-0,09	-0,18	-0,04	-0,02	0,01	0,08	0,16	0,46	0,25	0,27	0,32	0,76	0,77		0,75	0,66	0,46
	Sys4	0,55	0,59	0,62	0,71	0,53	0,44	0,05	-0,01	-0,10	-0,08	0,00	0,06	0,16	0,12	0,31	0,53	0,18	0,29	0,73	0,70	0,75		0,57	0,46
	Sys5	0,66	0,68	0,58	0,54	0,71	0,67	0,17	0,05	0,04	0,12	0,12	0,21	0,29	0,27	0,35	0,29	0,63	0,50	0,75	0,72	0,66	0,57		0,79
NEAR	ROVER	0,52	0,52	0,43	0,43	0,53	0,67	0,30	0,18	0,13	0,17	0,21	0,25	0,40	0,39	0,38	0,37	0,58	0,70	0,60	0,56	0,46	0,46	0,79	

FIGURE 2 – Matrice de corrélation des mesures WER, ATENE, In et Near pour les 5 systèmes de RAP et le ROVER sur Liste-100

Suppressions				
Atene	Mot	WER	In	Near
1	à	7	16	57
873	il	**1**	9	7
9	de	5	**1**	**1**

Insertions				
Atene	Mot	WER	In	Near
10	dix	673	-	2220
8112	et	**14**	16	**4**
8184	des	87	**12**	23

Substitutions					
Atene	Ref	Hyp	WER	In	Near
2	de	deux	124	98	48
4863	il	qui	**38**	-	-
8	deux	de	116	**4**	**18**

TABLE 2 – Extrait des listes générées, avec le rang associé à chacune des métriques, Atene, WER, les erreurs dans les entités (In) et les erreurs autour des entités (Near). En gras apparaît le rang dans chaque liste de la première erreur de chaque type.

suppression se trouve sur un rang relativement bas pour les deux autres listes. L'importance de cette erreur ne ressort pas particulièrement pour les listes In et Near (respectivement rang 16 et 57).

L'insertion la plus importante relevée par Atene est celle du nombre *dix*. Ce mot se retrouve, dans le corpus de référence, toujours dans une entité (Montant, Date ou Heure). Les trois autres métriques ne la considèrent pas importante (rang 673 pour le WER, inexistant pour In et 2220 pour Near).

La première substitution, le *de* (préposition ou partitif) substitué par *deux* (nombre) présente un peu le même problème avec les listes fondées sur les métriques traditionnelles. Cette substitution présente un risque élevé de soit provoquer une fausse alarme (détection erronée d'une entité) soit de segmenter une entité longue. Mais elle n'est pas considérée comme importante par les autres métriques (rang 124 pour le WER, 98 pour In et 48 pour Near). À l'inverse la substitution inverse de *deux* vers *de* est considérée comme plus importante par les méthodes In et Near (rang 4 et 18).

La suppression la plus importante pour le WER est celle du pronom *il* et les insertions les plus

importantes sont celles de la conjonction de coordination *et* et du déterminant *des*. Ces erreurs ne semblent pas pouvoir avoir un impact fort sur la REN, où qu'elles se produisent. Mais elles sont fréquentes, probablement car ces mots sont monophoniques, et sont du coup considérées haut placées.

La supression de *de* est considérée comme importante quelle que soit la méthode utilisée pour générer la liste d'erreurs. Ceci n'a rien d'étonnant puisque il s'agit d'une suppression fréquente et qu'en plus ce mot peut servir de marqueur pour la REN.

4 Conclusions et perspectives

Nous avons présenté une méthode pour générer des listes d'erreurs produites par une système de RAP ordonnées selon leur impact sur une tâche de détection d'entités nommées. Cette méthode s'appuie d'une part sur l'alignement tel que généré par le calcul de la métrique WER pour générer la liste initiale et la méthodologie liée à la métrique ATENE pour l'estimation de l'impact des erreurs.

Nous avons appliqué cette méthode aux données de test de la campagne ETAPE et avons comparé la liste ordonnée telle que générée par notre méthode avec des listes générées par d'autres méthodes. Les mesures de corrélation montrent que les métriques WER et ATENE sont cohérentes pour chaque système et fournissent une information en grande partie indépendante du système. Elles sont de plus assez peu corrélées entre elles, donnant des informations différentes sur l'importance des erreurs.

Une analyse plus détaillée des listes a ensuite permis de montrer que la quantification d'impact via ATENE fournit bien une information importante pour la tâche REN et pas uniquement fréquencielle. Nous avons donc bien réussi a établir une liste ordonnée d'erreurs motivée par le système aval.

L'étape suivante sera bien évidemment d'exploiter cette liste. Corriger le système de RAP n'est probablement pas possible, les erreurs importantes sont en l'occurence des erreurs difficiles. Nous essairons donc de prendre le problème dans l'autre sens et de rendre le système REN plus robuste face aux erreurs identifiées.

Remerciements

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

Références

ADDA-DECKER M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. In *Proc of JEP*, Dinard, France.

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

BOHÁČ M., NOUZA J. & BLAVKA K. (2012). Investigation on most frequent errors in large-scale speech recognition applications. In *Text, Speech and Dialogue*, p. 520–527 : Springer.

COMAS P. R. & TURMO J. (2009). Robust question answering for speech transcripts : Upc experience in qast 2009. In *Working Notes of CLEF 2009*, Corfou, Grèce.

DINARELLI M. & ROSSET S. (2011). Models Cascade for Tree-Structured Named Entity Detection. In *IJCNLP*, p. 1269–1278, Chiang Mai, Thailand.

DUFOUR R. & ESTEVE Y. (2008). Correcting asr outputs : Specific solutions to specific errors in french. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, p. 213–216.

GALIBERT O., LEIXA J., ADDA G., CHOUKRI K. & GRAVIER G. (2014). The ETAPE speech processing evaluation. In *LREC*, Reykjavik, Iceland.

GAROFOLO J. S., AUZANNE C. G. & VOORHEES E. M. (2000). The trec spoken document retrieval track : A success story. *NIST SPECIAL PUBLICATION SP*, **500**(246), 107–130.

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.

GHANNAY S., ESTÈVE Y. & CAMELIN N. (2015). Word embeddings combination and neural networks for robustness in asr error detection. In *EUSIPCO*, Nice, France.

GORYAINOVA M., GROUIN C., ROSSET S. & VASILESCU I. (2014). Morpho-syntactic study of errors from speech recognition system. In *LREC*, Reykjavik, Iceland.

HE X., DENG L. & ACERO A. (2011). Why word error rate is not a good metric for speech recognizer training for the speech translation task ? In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, p. 5632–5635 : IEEE.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, **10**(8), 707–710.

LIPPMANN R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, **22**(1), 1–15.

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.

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*.

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

RENA NEMOTO I. V. & ADDA-DECKER M. (2008). Speech errors on frequently observed homophones in french : Perceptual evaluation vs automatic classification. In *LREC*, Marrakech, Morocco.

SANTIAGO F., ADDA-DECKER M. & DUTREY C. (2015). Towards a typology of asr errors via syntax-prosody mapping. In *Errare Workshop*, Sinaia, Romania.

SCHARENBORG O. (2007). Reaching over the gap : A review of efforts to link human and automatic speech recognition research. *Speech Communication*, **49**(5), 336–347.

SPEARMAN C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, **15**, 72–101.

VASILESCU I., ADDA-DECKER M. & LAMEL L. (2012). Cross-lingual studies of ASR errors : paradigms for perceptual evaluations. In *LREC*, Istanbul, Turkey.

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.