Détection et Caractérisation d'erreurs dans des transcriptions automatiques pour des systèmes de traduction parole-parole

Frédéric Béchet, Benoit Favre
LIF/CNRS, Aix-Marseille Université, 163 avenue de Luminy, Marseille, France
prenom.nom@lif.univ-mrs.fr

Résumé

Quelle que soit la qualité des modules de transcription de la parole, les erreurs de reconnaissance sont inévitables étant donné d'une part les ambiguïtés intrinsèque à toute langue naturelle et d'autre part aux limites technologiques des systèmes actuels (vocabulaire fermé, séquentialité des traitements). L'étude et la caractérisation des erreurs de transcription représente un champ d'étude à part entière avec comme finalité d'atténuer l'impact de ces erreurs sur tout module exploitant des transcriptions automatiques. Dans le cadre d'un système de traduction parole-parole cette étude présente un module de détection et de caractérisation d'erreurs basé sur un étiqueteur de séquence utilisant des indices acoustiques, lexicaux et syntaxiques.

Abstract

Even though small ASR errors might not impact downstream processes that make use of the transcript, larger error segments like those generated by OOVs can have a considerable impact on applications such as speech-to-speech translation and can eventually lead to communication failure between users of the system. This work focuses on error detection in ASR output targeted towards significant error segments that can be recovered using a dialog system. We propose a CRF system trained to recognize error segments with ASR confidence-based, lexical and syntactic features.

Mots-clés : Traduction parole-parole, mesures de confiance, détection d'erreurs, analyse linguistique.

Keywords: Speech-to-speech translation, confidence measures, error detection, linguistic analysis.

1 Introduction

Les performances actuelles des systèmes de Reconnaissance Automatique de la Parole ont permis le développement de systèmes de communication innovants entre humains et machines tels que les assistants personnels commandés par la voix ou bien les traducteurs vocaux permettant à deux interlocuteurs de mener une conversation orale, chacun dans une langue différente.

Ces systèmes comportent tous un module de transcription automatique de la parole permettant de traduire le signal vocal en texte pouvant servir d'entrée au module d'interprétation sémantique, pour les systèmes de communication humain-machine, ou bien au module de traduction automatique pour les systèmes de traduction parole-parole. Cependant, quelle que soit la qualité du module de transcription, les erreurs de transcriptions sont inévitables étant donné d'une part les ambiguïtés intrinsèque à toute langue naturelle et d'autre part les limites technologiques des systèmes actuels (vocabulaire fermé, séquentialité des traitements).

Pouvoir détecter ces erreurs est crucial pour les traitements successifs des systèmes de transcription afin d'éviter de les propager au niveau de l'interprétation ou de la traduction. Cependant toutes les erreurs n'ont pas le même impact sur les performances globales d'un système. Ainsi une erreur d'accord, une confusion entre plusieurs formes conjuguées d'un même verbe ou bien l'ajout ou l'omission d'un mot grammatical tel que déterminant ou préposition, peuvent avoir un impact sur la structure syntaxique d'un énoncé, sans pour autant affecter de manière notable les performances. En effet les modules linguistiques utilisant les transcriptions automatiques sont entraînés à pouvoir traiter des énoncés bruités de ce type. En revanche, faire une confusion sur un syntagme prédicatif ou sur un groupe nominal peut avoir un impact très fort sur le système global.

Les mesures de confiance données par les systèmes de transcription automatique, telles que les probabilités a posteriori, ne prennent pas en compte cette notion d'impact. L'étude menée dans cet article consiste à proposer des indices supplémentaires permettant de mieux détecter et caractériser ces erreurs de transcription, dans le cadre d'un système de traduction parole-parole.

Le contexte expérimental de cette étude est celui de la tâche de traduction parole/parole du projet DARPA BOLT (Ayan et al., 2013). Cette tâche a pour but d'améliorer les systèmes de traduction parole/parole en permettant à la machine, en cas d'ambiguïté détectée lors de la transcription ou de la traduction, d'engager un dialogue de clarification avec l'utilisateur.

Cet article est organisé de la manière suivante : le paragraphe 2 présente des travaux connexes à cette étude dans le domaine de la détection d'erreurs et de mots hors-vocabulaire ; le paragraphe 3 présente la tâche visée dans le cadre du projet BOLT ; le paragraphe 4 détaille le modèle de détection d'erreurs développé qui est évalué dans le paragraphe 5.

2 Travaux connexes

Estimer la confiance d'une hypothèse de transcription pose plusieurs problèmes : choisir la granularité de l'erreur à détecter (Hazen et al., 2002) (mot, segment conceptuel, syntagme ou phrase entière), définir l'ensemble de paramètres qui seront intégrés dans le calcul de la

confiance (paramètres directement issus du processus de transcription, indices syntaxiques, indices contextuels), combiner les différents paramètres et mettre au point une stratégie de décision prenant en compte l'ensemble d'entre eux (Sarikaya et al., 2005). Quel que soit le modèle développé, la majorité des approches partage tout du moins deux étapes de base : générer autant de paramètres que possible pour caractériser chaque hypothèse produite ; estimer des scores de confiance à partir de ces paramètres ou indices.

Un problème connexe à celui de la détection d'erreurs de transcription est celui de la détection de segments de mots hors-vocabulaire. En effet, étant donné que tout système de transcription se base sur un vocabulaire fermé, tout mot inconnu de ce vocabulaire va générer nécessairement un segment d'erreur. Les systèmes actuels de détection de ces segments d'erreurs correspondant à des mots hors-vocabulaire sont basés sur des classifieurs prenant en compte un très grand nombre de paramètres provenant de plusieurs étapes de décodage (avec des unités lexicales et sous-lexicales) (Rastrow et al., 2009) ainsi que des indices prosodiques ou syntaxiques (Hirschberg et al., 2004). Ce problème est envisagé sous l'angle d'un problème de classification binaire ou chaque mot (ou segment d'un réseau de confusion) est classé comme mot hors-vocabulaire (Out Of Vocabulary - OOV) ou pas.

Plus récemment des études telles que (Parada et al., 2010) ont proposé de considérer le problème de la détection de segments hors-vocabulaire comme un problème d'étiquetage de séquence. L'intuition derrière cette idée est basée sur le fait que les mots hors vocabulaire ont tendance à générer plusieurs erreurs, d'une part en remplaçant chaque mot inconnu par une séquence de mots connus plus petits, et d'autre part à cause de l'effet "boule de neige" provoqué par ces même mots dans leur contexte immédiat. Une approche à base de Conditional Random Fields (CRF) a été proposée, intégrant tous les indices présentés auparavant, et visant à prédire la meilleure séquence d'étiquettes `B_OOV` (début de segment OOV), `I_OOV` (à l'intérieur d'un segment OOV) et `O_OOV` (en dehors d'un segment OOV) à partir d'une séquence de mots hypothèses. Cette approche a obtenu des gains significatifs en performance comparée à l'approche basée uniquement sur des décisions locales sur les mots.

Nous proposons dans cette étude de généraliser cette méthode à tous les différents types d'erreurs de reconnaissance. Cet article est une extension de (Béchet et Favre, 2013). En se basant sur une approche à base de CRF, nous générons un graphe d'hypothèses de segments d'erreurs permettant d'améliorer la détection de séquences d'erreurs ayant un impact important sur les performances des traitements suivants la phase de transcription de la parole.

3 Erreurs de reconnaissance et traduction parole-parole

Cette étude a été menée dans le cadre de la tâche de traduction parole-parole du projet DARPA BOLT. Le cadre applicatif est la communication entre deux personnes ne parlant pas la même langue, sans support visuel (uniquement à la voix), pour des tâches de collecte d'information. L'originalité de BOLT, par rapport à des projets similaires tels que TRANSTAC, consiste à doter les modules de transcription et de traduction d'une capacité à générer un dialogue de clarification avec l'utilisateur lorsqu'une erreur est détectée (Ayan et al., 2013).

La difficulté principale de ce type de système est d'engager des dialogues de clarification uniquement lorsque les erreurs détectées sont susceptibles de bloquer le dialogue entre les

locuteurs.

Un exemple est donné dans la figure 1.

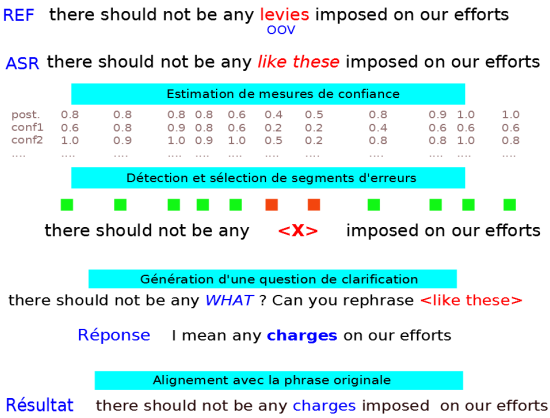

Figure 1 – Exemple de déroulement d'un dialogue de clarification dans BOLT

Cette différence d'importance entre erreurs de reconnaissance est illustrée par la table 1 présentant deux exemples provenant du corpus BOLT avec les hypothèses de transcription (ASR 1-best), les mesures de confiance du système de reconnaissance (ASR posteriors) et la séquence de mots effectivement prononcée par le locuteur (Réference).

Le premier exemple contient un mot hors-vocabulaire "acetaminophen" qui est remplacé par le système de reconnaissance par la séquence de 6 mots : "I see them in a fan". Dans ce cas, il est très important de détecter cette erreur et engager un dialogue de clarification avant la phase de traduction. Il est intéressant de constater qu'au niveau des mesures de confiance, seul le début et la fin du segment erroné ont des valeurs basses, les erreurs de milieu de segment ont toutes des valeurs acceptables. Ce phénomène justifie l'approche à base d'étiquetage de séquence utilisée dans cette étude.

exemple 1													
ASR 1-best	is	the	town	low	on	supplies	of	I	see	them	in	a	fan
ASR posteriors	1	1	1	0.9	1	1	1	0.1	0.9	0.9	0.9	0.8	0.6
Réference	is	the	town	low	on	supplies	of	acetaminophen					

exemple 2									
ASR 1-best	they	travel	to	cross	to	desert	on	camera	back
ASR posteriors	0.7	0.8	0.8	0.8	0.8	0.9	0.9	0.7	0.8
Réference	they	traveled		across	the	desert	on	camelback	

Table 1 – Exemple de transcriptions automatiques avec des segments d'erreurs

Le second exemple contient deux types d'erreurs : des confusions sur des mots proches ("travel/traveled", "cross/across"), une confusion sur un mot monosyllabique ("to/the"), enfin de nouveau une erreur due à un mot hors-vocabulaire ("camera back/camelback").

Sur ces erreurs, deux clarifications seraient nécessaires dans cet exemple avant d'effectuer la traduction : d'une part l'ambiguïté entre "to desert" (déserter) et "the desert" ; d'autre part la clarification sur le mot hors-vocabulaire "camelback".

Il est clair sur ces exemples que les seules mesures de confiance de type probabilités a posteriori ne permettent ni de détecter facilement ces erreurs, ni de les distinguer les unes des autres. C'est pour cela que nous proposons un traitement intégrant à la fois des indices linguistiques, mais aussi une méthode à base d'étiquetage de séquences.

4 Détection d'erreurs par étiquetage de séquence

La méthode de détection de segments d'erreurs présentée dans cette étude est basée sur un étiqueteur de séquence de type Conditional Random Field (CRF). Cet étiqueteur va attribuer un label binaire à chaque mot : erreur ou correct.

Réference

word	we	need	judicious	men
POS	PRP	VBP	JJ	NN
synt. dep.	SBJ(need)	ROOT()	NMOD(men)	OBJ(need)

ASR 1-best

word	we	need	your	dishes	that
POS	PRP	VBP	PRP	NN	WDT
synt. dep.	SBJ(need)	ROOT()	NMOD(dishes)	OBJ(need)	NONE()

Table 2 – Exemple d'analyse linguistique de transcription de référence et d'hypothèses

Cet étiqueteur est appris sur un corpus de transcriptions automatiques sur lesquelles les étiquettes erreur et correct sont obtenues par alignement dynamique avec les transcriptions de référence. Trois niveaux de paramètres sont utilisés pour décrire les exemples du corpus d'apprentissage :

1. Paramètres issus du système de transcription de parole : nous utilisons les probabilités a posteriori générées durant le décodage. Ces probabilités sont discrétisés par la méthode décrite dans (Fayyad et Irani, 1993) avec l'implémentation de (Raymond, 2012). Nous utilisons les valeurs du mot courant, du mot précédent et du mot courant.
2. Paramètres lexicaux : le mot courant et son contexte droit et gauche font partie des paramètres. Ils sont associés à un paramètre relatif à la taille du mot (en nombre de caractères) ainsi qu'à un paramètre indiquant si les 3 trigrammes contenant le mot courant et son contexte immédiat ont été vus dans le corpus d'apprentissage du modèle de langage.
3. Paramètres syntaxiques : les transcriptions sont analysées avec la boite à outils MACAON (Nasr et al., 2011). Nous utilisons un étiqueteur morphosyntaxique ainsi qu'un analyseur en dépendance. Les séquences d'étiquettes syntaxiques sont utilisées comme paramètres, ainsi que les liens de dépendance entre le mot courant et son gouverneur.

Un exemple d'analyse linguistique est donné dans la table 2. Tous les paramètres syntaxiques sont produits à partir de ces analyses.

5 Expérimentation et résultats

Nous utilisons dans cette étude des corpus en langue anglaise provenant du projet DARPA BOLT et transcrit automatiquement par le système de SRI Dynaspeak (Franco et al., 2002). Ce corpus contient deux types de phrases :

– Corpus 1 : cette partie du corpus contient des dialogues enregistrés durant le projet TRANSTAC, contenant des conversations centrées sur la collecte d'information, proche des corpus utilisés pour entraîner le système de reconnaissance. Il y a peu de mots hors-vocabulaire et le taux d'erreurs mots est assez faible ($< 10\%$).
– Corpus 2 : ce corpus a été défini durant le projet BOLT pour contenir spécifiquement des phrases difficiles avec des ambiguïtés provenant soit d'un mot hors-vocabulaire, soit d'une ambiguïté de type homophone, soit d'un mot volontairement mal prononcé. Le taux d'erreur de ce corpus est bien plus élevé (35%).

Ces corpus correspondent ainsi à deux situations différentes : le premier corpus représente les erreurs classiques d'un système de reconnaissance, qui peuvent évidemment perturber le système global, mais qui ne nécessitent pas nécessairement un dialogue de clarification. Le deuxième corpus contient les erreurs pouvant avoir un impact important sur la suite des traitements qui doivent être clarifiées avant de pouvoir lancer la phase de traduction.

Corpus	#mots	#phrases	WER	taille err.	fertilité
Corpus 1	84405	6527	8.4	1.5	1.2
Corpus 2	4919	570	35.8	2.6	4.8

Table 3 – Description des corpus avec leurs caractéristiques

Ces deux corpus sont décrits dans la table 3. Nous remarquons que la taille moyenne des erreurs est bien supérieure pour le corpus 2. La fertilité d'une erreur correspond aux nombres de mots erronés générés par le système de transcription pour un mot mal reconnu dans la transcription de référence. La fertilité moyenne donnée dans le tableau indique ainsi le nombre moyen d'insertion observé dans les transcriptions automatique pour un mot de la référence non reconnu. Nous remarquons que pour le corpus 2, la fertilité est très forte : presque 5 insertions pour chaque erreur. Ceci s'explique par le fait que les mots hors-vocabulaire ont tendance à être remplacés par des séquences de mots courts, comme illustré dans la table 1. Afin de ne pas faire d'a priori sur le type d'erreurs que l'on va rencontrer, les résultats présentés ici ont été obtenus avec une approche de type 10-fold, où les corpus 1 et 2 ont été mélangés, puis découpés en 10 partitions. L'étiqueteur CRF (Okazaki, 2007) est appris sur les 9/10 de chaque partition, puis évalué sur la dixième. Il y a plusieurs méthodes pour évaluer les performances d'un détecteur d'erreurs. La méthode la plus simple consiste à évaluer la prédiction erreur/correct pour chaque mot. Cette métrique est présentée dans la table 4 dans la colonne correct. Etant donné que la plupart des mots sont corrects, cette mesure n'est pas très informative. Il est plus intéressant de se placer dans un contexte de détection d'évènements où les seules évènements à détecter sont les erreurs de reconnaissance et non pas les mots corrects. Nous donnons ainsi les résultats en termes de précision (P), rappel

(R), Fausse ALarme (FA), et détection manquée (Miss) pour les erreurs de reconnaissance uniquement (un système prédisant que tous les mots sont corrects aurait 0% de précision et rappel). La table 4 présente ces résultats séparément pour les deux corpus, bien qu'ils aient été réunis durant la phase d'apprentissage et de test. Plusieurs conditions sont comparées : en utilisant dans le modèle CRF uniquement les probabilités a posteriori, puis en rajoutant les paramètres lexicaux, et enfin les paramètres syntaxiques.

Corpus 1					
Paramètres	correct	P	R	FA	Miss
ASR post.	94.8	51.1	22.0	48.9	78.0
+ lexical	95.1	55.1	32.1	44.9	67.9
+ syntaxique	95.1	55.4	36.8	44.6	63.2
Corpus 2					
Paramètres	correct	P	R	FA	Miss
ASR post.	78.7	76.6	41.5	23.5	58.5
+ lexical	81.3	74.0	57.8	26.0	42.2
+ syntaxique	82.4	74.0	63.7	26.0	36.3

Table 4 – Résultats en terme de classification erreur/non erreur (correct) ; puis de détection des erreurs uniquement sur les deux corpus C1 et C2

Comme on peut le voir les prédictions d'erreurs sont meilleures sur Corpus 2. C'est d'une part attendu, car les erreurs sont plus saillantes, mais d'autre part c'est aussi espéré car il s'agit des erreurs pour lesquelles il faut avoir une stratégie de correction. Les courbes (Miss/FA) sont présentées dans la figure 2. Elles sont obtenues en faisant varier un seuil sur le score de détection d'une erreur donné par l'étiqueteur CRF.

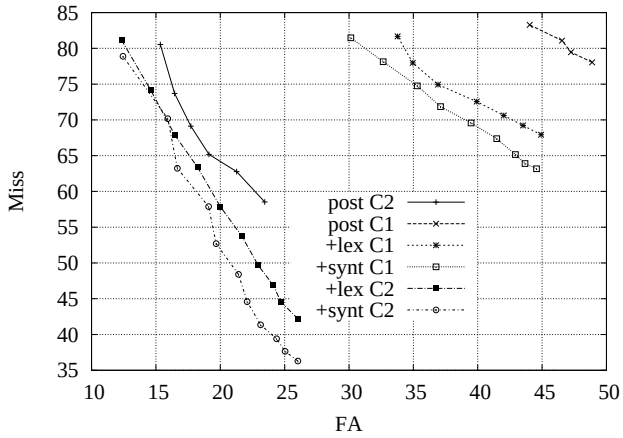

Figure 2 – Courbe Miss/FA avec différents jeu de paramètres sur les corpus C1 et C2

Toutes les mesures précédentes sont données au niveau des mots. Cependant, dans cette étude, on s'intéresse principalement aux segments d'erreurs, considérant la fertilité forte des erreurs impactant les performances finales du système de traduction. Nous présentons ainsi une mesure plus proche de la tache visée : la mesure du taux d'erreurs mots après avoir agrégé toutes les détections successives d'erreurs au sein d'un même symbole `X`. Cette méthode ne permet pas de corriger des erreurs de transcription, elle permet cependant de diminuer le nombre d'insertions erronées. Notons aussi que c'est une mesure assez dure dans la mesure où aucun gain n'est donné pour la détection correcte d'une erreur constituée d'un seul mot : un mot erroné est remplacé par `X`, qui est toujours considéré comme une confusion par le programme d'alignement avec la référence. Par exemple, sur une phrase comme "I saw that man at Izamni", si la transcription automatique est "I saw that man at is on me" on aura un taux d'erreurs de 50%. Si on détecte correctement le segment d'erreurs et qu'il est remplacé dans la transcription automatique par le symbole `X`, nous obtenons la transcription : "I saw that man at X" qui a un taux d'erreurs de 16.6%. Cette mesure alternative de la qualité d'un prédicteur d'erreurs de reconnaissance est particulièrement intéressante car elle permet de juger de l'utilité objective d'un tel prédicteur (baisse du taux d'erreurs sur les mots) sans être liée à un cadre applicatif précis.

Corpus/WER	ASR	Oracle	P>0.5	P>0.8
Corpus 1	8.4	7.5	9.2	8.3
Corpus 2	35.8	19.9	31.1	29.6

Table 5 – Taux d'erreur mots obtenus en agrégeant tous les segments d'erreurs détectés automatiquement dans les transcriptions automatiques par un seul symbole `X` puis en comparant à la référence

On peut voir dans la figure 5, grâce au taux Oracle, l'importance de bien détecter les segments d'erreurs. On peut faire chuter très significativement le taux d'erreurs mots, même sans la moindre correction, si on arrive à détecter correctement les segments erronés. Les taux d'erreurs sont donnés selon deux conditions : P>0.5 (respectivement P>0.8) signifie qu'un mot est considéré comme erreur si son score de confiance est supérieur à 0.5 (respectivement 0.8). On voit qu'une baisse sensible du taux d'erreurs mots sur les corpus 2 est constaté (gain de 6% absolu). Ce résultat est encourageant, il montre que notre approche basée sur un étiqueteur de séquence et des indices linguistiques parvient bien à détecter les segments d'erreurs importants, sans pour autant impacter les performances pour le corpus 1.

6 Conclusion

Dans le cadre d'un système de traduction parole-parole cette étude a présenté un module de détection et de caractérisation d'erreurs basé sur un étiqueteur de séquence utilisant des indices acoustiques, lexicaux et syntaxiques. La partie expérimentale, effectuée sur le corpus BOLT, a montré la validité de l'approche pour la détection des erreurs "importantes" correspondant par exemple aux mots hors vocabulaire, qui ont un impact important sur les traitements successifs à la phase de reconnaissance.

Remerciements

Cette étude a été financée en partie par le contrat BOLT DARPA HR0011-12-C-0016 en tant que sous-traitant de SRI International.

Références

Ayan, N. F., Mandal, A., Frandsen, M., Zheng, J., Blasco, P., Kathol, A., Béchet, F., Favre, B., Marin, A., Kwiatkowski, T. et al. (2013). "can you give me another word for hyperbaric ?" : Improving speech translation using targeted clarification questions. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 8391–8395. IEEE.

Béchet, F. et Favre, B. (2013). Asr error segment localization for spoken recovery strategy. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 6837–6841. IEEE.

Fayyad, U. et Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning.

Franco, H., Zheng, J., Butzberger, J., Cesari, F., Frandsen, M., Arnold, J., Gadde, V., Stolcke, A. et Abrash, V. (2002). Dynaspeak : Sri's scalable speech recognizer for embedded and mobile systems. In Proceedings of the second international conference on Human Language Technology Research, pages 25–30. Morgan Kaufmann Publishers Inc.

Hazen, T., Seneff, S. et Polifroni, J. (2002). Recognition confidence scoring and its use in speech understanding systems. Computer Speech & Language, 16(1):49–67.

Hirschberg, J., Litman, D. et Swerts, M. (2004). Prosodic and other cues to speech recognition failures. Speech Communication, 43(1):155–175.

Nasr, A., Béchet, F., Rey, J., Favre, B. et Le Roux, J. (2011). Macaon : An nlp tool suite for processing word lattices. Proceedings of the ACL 2011 System Demonstration, pages 86–91.

Okazaki, N. (2007). Crfsuite : a fast implementation of conditional random fields (crfs) `http://www.chokkan.org/software/crfsuite/`.

Parada, C., Dredze, M., Filimonov, D. et Jelinek, F. (2010). Contextual information improves oov detection in speech. In North American Chapter of the Association for Computational Linguistics (NAACL).

Rastrow, A., Sethy, A. et Ramabhadran, B. (2009). A new method for oov detection using hybrid word/fragment system. In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, pages 3953–3956. IEEE.

Raymond, C. (2012). Discretized for CRF. `http://www.irisa.fr/texmex/people/raymond/Tools/tools.html`.

Sarikaya, R., Gao, Y., Picheny, M. et Erdogan, H. (2005). Semantic confidence measurement for spoken dialog systems. Speech and Audio Processing, IEEE Transactions on, 13(4):534–545.