Détection et caractérisation des régions d'erreurs dans des transcriptions de contenus multimédia : application à la recherche des noms de personnes

Richard Dufour, Géraldine Damnati, Delphine Charlet

Détection et caractérisation des régions d'erreurs dans des transcriptions de contenus multimédia : application à la recherche des noms de personnes

Richard Dufour Géraldine Damnati Delphine Charlet
France Telecom R&D - Orange Labs, 2, av. Pierre Marzin 22307 Lannion
prénom.nom@orange.com

RÉSUMÉ

Dans cet article, nous proposons de détecter et de caractériser des régions d'erreurs dans des transcriptions automatiques de contenus multimédia. La détection et la caractérisation simultanée des régions d'erreurs peut être vue comme une tâche d'étiquetage de séquences pour laquelle nous comparons des approches séquentielles (segmentation puis classification) et une approche intégrée. Nous comparons les performances de notre système sur deux corpus différents en faisant varier les données d'apprentissage. Nous nous intéressons particulièrement aux erreurs des noms de personnes, information essentielle dans de nombreuses applications d'extraction d'information. Les résultats obtenus confirment l'intérêt d'une méthode à base d'apprentissage exploitant le contexte d'apparition des erreurs.

ABSTRACT

Error region detection and characterization in transcriptions of multimedia documents : application to person name search

In this article, we propose to detect and characterize error regions in automatic transcriptions of multimedia documents. The simultaneous detection and characterization could be seen as a sequence labeling task where we compare sequential approaches (segmentation then classification) and an integrated one. We compare our system performance on two different corpus by varying training data. We are particularly interested in person name errors, essential information in various information extraction applications. Results confirm the interest for learning-based method using the apparition context of errors.

MOTS-CLÉS : régions d'erreurs, caractérisation des erreurs, transcription automatique, classification automatique, noms propres.

KEYWORDS: error regions, error characterization, automatic transcription, automatic classification, person names.

1 Introduction

Dans le cadre de la reconnaissance de la parole continue à grand vocabulaire, les systèmes de reconnaissance automatique de la parole (RAP) peuvent actuellement fournir des transcriptions avec un bon niveau de performance, permettant leur intégration dans de nombreuses applications. Cependant, les erreurs de transcription de ces systèmes sont inévitables, ce qui représente toujours un problème pour certains domaines, tel que l'extraction automatique d'information dans des

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 811–818,
Grenoble, 4 au 8 juin 2012.

documents multimédia. Dans cet article, l'objectif est d'identifier et de caractériser les erreurs dans les transcriptions automatiques. Pour ce faire, nous ne considérons pas simplement ces erreurs de manière isolée mais nous cherchons à détecter et caractériser des régions d'erreurs (*i.e.* des regroupements d'erreurs consécutives).

Traditionnellement, la détection d'erreurs est conduite au travers de la définition des mesures de confiance (MC) représentant la probabilité qu'un mot soit correct. Appliquer un seuil sur ce score permet au système d'être réglé à un point de fonctionnement donné pouvant être choisi en fonction du contexte d'application (choix entre rappel ou précision élevés). Les MC peuvent être vues comme des classifieurs binaires permettant de séparer les mots en corrects/incorrects, leur performance est généralement évaluée en fonction de leur capacité à retrouver les mots corrects. Cependant, lorsque le taux d'erreur-mots est bas, cette tâche de classification binaire est typiquement un problème de classification avec données déséquilibrées : l'évaluation centrée sur la classe majoritaire (mots corrects) masque la capacité du classifieur à gérer la classe minoritaire (mots mal transcrits). Dans cet article, nous nous intéressons à la détection des erreurs de transcription dans le cadre d'émissions d'information multimédia, avec, pour entrée, des jeux de données déséquilibrés (en faveur des mots corrects). Nous nous focaliserons sur l'évaluation de la capacité de notre système à correctement détecter les erreurs de reconnaissance.

Au delà de le détection des erreurs nous voulons également les caractériser afin de déterminer leur nature. En fait, toutes les causes d'erreurs n'ont pas le même impact selon le contexte d'application considéré. On peut décider d'ignorer une erreur si son impact est jugé négligeable ou d'éventuellement définir des stratégies de correction appropriées dans le cas contraire. D'un point de vue analytique, plusieurs études ont fourni une analyse détaillée *a posteriori* des causes des erreurs. Les auteurs dans (Duta *et al.*, 2006) ont mis en lumière le fait que la majorité des erreurs de transcription dans les émissions d'information en langue anglaise sont dues à des entités nommées. Dans (Vasilescu *et al.*, 2009), les auteurs ont montré que les homophones, dans la langue française, sont très fréquents et représentent une importante source d'erreurs pour les systèmes de RAP. Cependant, du point de vue de leur caractérisation automatique, de nombreuses études se sont focalisées sur la détection et la correction des mots hors-vocabulaires (HV), dont le comportement et l'impact diffèrent des autres erreurs (Woodland *et al.*, 2000). Des stratégies spécifiques ont été proposées pour détecter les mots HV en utilisant, par exemple, un modèle de langage hybride de mots et de sous-mots (Rastrow *et al.*, 2009). Dans (Parada *et al.*, 2010), les auteurs se sont intéressés aux régions d'erreurs générées par les mots HV et ont proposé une méthode prenant en compte l'information contextuelle des régions voisines au lieu de ne considérer que la région "locale" des mots HV. Leur corpus a été construit en ne conservant que les mots HV riches en sens, en excluant ceux ayant moins de 4 phonèmes, et en supposant que les frontières des régions sont connues à l'avance. Dans les langues fortement flexionnelles, les mots HV peuvent être de natures différentes. Ainsi, bien que les noms propres impliqués dans les entités nommées soient une source importante des mots HV, d'autres causes sont à considérer, telles que les flexions d'un lemme donné, ou encore la présence de mots rarement utilisés dans le langage courant. Réciproquement, il arrive que des noms propres pourtant présents dans le dictionnaire soient mal transcrits. Ainsi, nous avons choisi de ne pas nous focaliser sur les mots HV, mais de définir des classes plus pertinentes pour notre contexte d'application en traitant la nature des erreurs dans leur ensemble (mot HV ou non). Bien que nous considérions toutes les erreurs, nous nous intéressons plus particulièrement aux noms de personnes, dont les erreurs ont un impact fort dans de nombreuses applications.

Dans cet article, nous traitons la détection et la caractérisation des régions d'erreurs comme une

tâche d'étiquetage de séquences. Nous cherchons à comprendre l'impact du contexte d'apparition des erreurs pour cette tâche particulière. Nous détaillerons les données expérimentales (partie 2) puis les approches d'étiquetage de séquences décrites dans (Dufour *et al.*, 2012) ainsi que leur évaluation (parties 3 et 4). Enfin, les expériences menées sur l'influence de la base d'apprentissage seront présentées dans la partie 5, en analysant particulièrement les noms de personnes.

2 Données expérimentales

2.1 Description des données

Les expériences que nous menons s'appuient sur deux corpus expérimentaux en langue française manuellement transcrits et dont les noms de personnes ont été annotés. Le premier corpus est composé de 38 journaux télévisés (information, interviews, reportages...) collectés à partir de 7 chaînes de télévision entre octobre 2008 et janvier 2009. Ce corpus a ensuite été découpé en deux ensembles, avec 24 émissions pour (*JT train*) et 14 émissions pour (*JT test*). Le second corpus est composé de 28 extraits d'émissions télévisées plus hétérogènes (journaux télévisés, débats, émissions culturelles) provenant du corpus de développement du défi *REPERE* [1]. Il n'y a pas de recouvrement entre les chaînes dont sont extraites les données *JT* et celles des données *REPERE*. Les transcriptions automatiques de ces émissions sont réalisées au moyen du système de reconnaissance de la parole *VoxSigma v3.5* de *Vocapia Research* et fondé sur la technologie développée au LIMSI (Gauvain *et al.*, 2002). Les différents corpus sont décrits dans le tableau 1.

TABLE 1 – Description des corpus *JT train*, *JT test* et *REPERE*

	JT train	**JT test**	**REPERE**
Durée	7h45	6h15	3h00
Nb mots (taux d'erreur-mots)	84 146 (15,9 %)	70 538 (18,0 %)	33 413 (21,2 %)
Nb régions d'erreurs (taille moyenne)	5 529 (1,8)	4 908 (1,8)	2 296 (2,1)

Nous utilisons les mesures de confiance des mots estimées par le système de transcription (probabilités *a posteriori* calculées à partir des graphes de mots). Cette mesure de confiance est performante, avec un score d'entropie croisée normalisée respectivement de 0,36 sur le corpus complet des *JT* (train et test) et de 0,31 sur le corpus *REPERE*. Ces mesures de confiance sont notamment utilisées par le système de transcription pour filtrer les hypothèses émises : par défaut, les mots ayant une mesure de confiance inférieure à 0,3 sont retirés des hypothèses de la transcription. Cette étape permet d'améliorer le taux d'erreur-mots final. Ainsi, pour le corpus *JT (train+test)* le taux d'erreur-mots avec les mots filtrés serait de 19,6 % (principalement à cause des insertions de mots) et sur le corpus *REPERE* de 24,3 %.

2.2 Définition des classes d'erreurs

Le système que nous avons développé cherche à identifier 4 sources d'erreurs déterminées à partir de l'alignement entre les transcriptions automatiques et manuelles. L'alignement a été réalisé au moyen de l'outil NIST Sclite [2]. En premier lieu, nous avons défini la classe *Nom de personne (NP)*, particulièrement étudiée dans cette article, puisque cette information est essentielle dans de nombreuses applications d'extraction d'information. Nous avons également défini la classe *Autre nom propre (ANP)*, pouvant également contenir des informations très utiles. La classe *Homophone (H)* [3] a été choisie afin de prendre en compte un phénomène très fréquent dans la

1. `http://www.defi-repere.fr`
2. `http://www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm`
3. Comparaison des empreintes phonétiques des mots hypothèses et de référence au moyen d'un lexique additionnel.

langue française mais qui est moins pénalisant dans une optique d'indexation. Enfin, les erreurs ne rentrant dans aucune de ces classes ont été regroupées au sein de la classe *Autre (A)*. Lorsque plusieurs causes peuvent être attribuées à une même région, un ordre de priorité a été défini : **1 -** *Nom de personne (NP)*, **2 -** *Autre nom propre (ANP)*, **3 -** *Homophone (H)* et **4 -** *Autre (A)*. À titre d'exemple, les régions de type *A* dans le corpus *REPERE* représentent 68 % des régions d'erreurs (8 % pour *NP*, 4 % pour *ANP* et 20 % pour *H*). Si la classe des NP représente une faible proportion des régions d'erreurs, elle reste d'une importance applicative particulière d'autant que 37,4 % des entités NP prononcées initialement génèrent une région d'erreurs. Des tendances assez proches se dessinent sur les deux corpus, à savoir que les *NP* et les *ANP* génèrent des régions d'erreurs de tailles plus grandes que les erreurs dues à des *H* ou *A* (2,5 erreurs consécutives en moyenne pour les *NP* sur le corpus *REPERE* et 1,6 pour les *H*).

3 Extraction et caractérisation des régions d'erreurs

La détection et la caractérisation simultanée des régions d'erreurs peut être vue comme une tâche d'étiquetage de séquences. Nous proposons dans la sous-partie 3.1 une approche séquentielle qui consistera, dans un premier temps, à segmenter les transcriptions en région correcte / erronée, et dans un second temps à associer une classe à ces régions d'erreurs. Puis nous proposons une approche intégrée en 3.2 consistant à segmenter et étiqueter conjointement en classes d'erreurs. De plus amples détails peuvent être trouvés dans (Dufour *et al.*, 2012).

3.1 Approche séquentielle

Nous proposons trois approches différentes pour segmenter en régions d'erreurs. En premier lieu, nous utilisons une approche classique *Base* consistant à appliquer un seuil θ_b sur les mesures de confiance fournies par le système de RAP. En fait, les mots consécutifs dont le score est inférieur à θ_b seront considérés comme une région d'erreurs.

Appliquer un seul seuil sur les mesures de confiance peut ne pas être suffisant puisque les erreurs consécutives ne sont pas toutes associées à une mesure de confiance basse. Afin d'assouplir cette contrainte, nous introduisons un automate à deux états. Chaque mot d'un segment est analysé : dans l'état *Correct*, le mot est considéré comme correctement transcrit, alors que l'état *Erreur* détecte le mot comme incorrect. Le seuil θ_{err} permet de passer de l'état *Correct* à *Erreur*, ou de rester dans l'état *Correct*, et inversement pour le seuil θ_{cor} (avec $\theta_{err} < \theta_{cor}$). L'automate sera utilisé pour chaque phrase dans les deux sens de lecture (de droite à gauche et vice versa) afin de capturer les erreurs non trouvées dans un sens. Des automates d'ordre plus élevé ont été implémentés mais fournissaient des régions d'erreurs beaucoup trop grandes. Dans cette approche, nous ne nous intéressons pas simplement à la MC courante, mais à celles se trouvant au voisinage. Par exemple, il est possible de rester dans l'état *Erreur* si la mesure de confiance est située entre θ_{err} and θ_{cor}.

Enfin, nous proposons d'utiliser les champs conditionnels aléatoires (CRF) (Lafferty *et al.*, 2001) pour délimiter les régions d'erreurs en prenant en compte de nombreuses sources d'information : les bigrammes de mots, l'étiquetage grammatical et regroupement en syntagmes [4], les mesures de confiance, et les durées du mot courant, précédent et suivant. La mise en œuvre repose sur un formalisme *UIO* (*Unique* pour les erreurs isolées, *Inside* pour les $n > 1$ erreurs consécutives et *Outside* pour les mots corrects). *Begin* n'est pas pris en compte car peu performant.

Après avoir détecté ces régions d'erreurs, nous proposons de les associer à une des quatre classes d'erreurs décrites dans la section 2.2 au moyen d'une méthode de classification. Nous avons

4. Lia_tagg : `http://pageperso.lif.univ-mrs.fr/~frederic.bechet`

choisi d'utiliser l'outil *Icsiboost* [5], un classifieur à larges marges fondé sur l'algorithme *AdaBoost*. De nombreuses caractéristiques sont utilisées : les mots des régions (bigrammes), l'étiquetage grammatical et regroupement en syntagmes (trigrammes), le nombre des mots de la région, quadrigrammes sur les cinq mots précédents, la durée et la moyenne des mesures de confiance de chaque tour de parole, et enfin le nombre de syllabes par mot.

3.2 Méthode intégrée

Comme les CRF peuvent segmenter et étiqueter des séquences de données, nous proposons d'utiliser cette méthode pour directement retrouver les régions d'erreurs et les étiqueter avec une des 4 classes d'erreurs au lieu de réaliser ces opérations séparément. Les mêmes caractéristiques que celles présentées pour l'approche utilisant les CRF de manière séquentielle seront utilisées. Le formalisme *UIO* est toujours utilisé, auquel on associe les 4 classes d'erreurs. Au final, cela revient à utiliser 9 classes ($4 * I + 4 * U + O$).

Nous proposons une dernière solution consistant à combiner toutes les propositions en fusionnant les régions d'erreurs au moyen de l'opérateur "OU", et en choisissant ensuite la classe de la région d'erreurs en fonction de la priorité définie dans la partie 2.2.

4 Évaluation comparée des différentes approches

4.1 Optimisation des seuils de décision

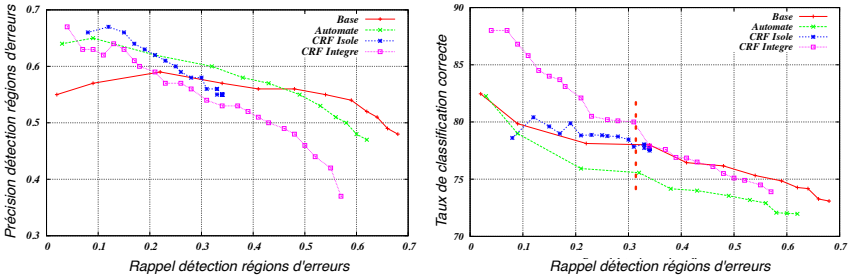

FIGURE 1 – Précision de la segmentation (à gauche) et taux de classification correcte des régions bien segmentées (à droite) en fonction du rappel de la détection sur le corpus *JT train*

Afin d'évaluer la performance des différentes approches, nous présentons dans la figure 1 (à gauche) les résultats obtenus sur le corpus *JT train* en termes de rappel et de précision sur la détection des régions d'erreurs en faisant varier le seuil de décision des 4 approches proposées. Détecter précisément les régions d'erreurs étant une tâche difficile, nous avons choisi d'assouplir la détection des régions en considérant comme correcte une région d'erreurs dont les frontières sont erronées à deux mots près. Les méthodes *Base* et *Automate* permettent d'atteindre des taux de précision plus élevés que les méthodes à base de CRF lorsque le rappel est très haut, mais à l'inverse, les méthodes à base de CRF sont plus précises lorsque le rappel est faible. Cette plus faible précision pour les CRF s'explique par un nombre trop grand d'hypothèses de détection conduisant à de nombreuses insertions ainsi qu'à des régions trop longues. Ces difficultés sont mieux gérées avec l'utilisation de la mesure de confiance seule. Notons également

5. `http://code.google.com/p/icsiboost`

que le comportement des deux approches à base de CRF diffère au niveau de l'évolution du taux de rappel : l'approche *CRF Isolé* ne dépasse pas les 35 % en rappel alors que la variation du seuil de décision permet à l'approche *CRF Intégré* d'approcher les 60 %. Il apparaît donc que, du point de vue de la segmentation seule en régions d'erreurs, le phénomène est mieux modélisé lorsque les classes sont utilisées dans le processus de segmentation (*CRF Intégré*, 9 classes) plutôt qu'une détection globale des régions d'erreurs (*CRF Isolé*, 3 classes).

Pour définir le seuil de décision, nous nous intéressons au taux de bonne classification des régions correctement détectées. Notre choix se porte sur un seuil optimisé en fonction du taux de rappel de la détection et de la qualité de la classification de ces régions (figure 1 à droite). Nous avons choisi de nous placer à un taux de rappel à peu près équivalent pour chaque approche autour de 0,3 afin de ne pas perturber la fusion finale.

À ce point de fonctionnement, la *fusion* des approches conduit à un taux de rappel de 42,2 % (gain de 8 points en absolu par rapport à la meilleure méthode) en gardant une précision acceptable de 57,0 %. Les régions bien détectées conduisent à un taux de classification correcte de 78,4 %.

4.2 Évaluation globale

Les performances de la détection des régions d'erreurs sont évaluées en rappel/précision. Le taux de classification correcte évalue la classification des classes d'erreurs sur les régions correctement détectées. Par ailleurs, une nouvelle mesure inspirée du Slot Error-Rate (SER) (Makhoul *et al.*, 1999) est introduite afin d'évaluer les performances globales de la détection et caractérisation des régions d'erreurs. Cette métrique est particulièrement utilisée pour évaluer les systèmes de détection des entités nommées. Elle possède l'avantage de prendre en compte de nombreuses combinaisons d'erreurs potentielles contenues dans notre double problématique de détection et caractérisation de régions d'erreurs :

$$SER = \frac{D + I + S_{all} + 0,5 * (S_{cla} + S_{reg})}{Nombre\ total\ des\ régions\ d'erreurs\ de\ référence} \quad (1)$$

où D est le nombre de régions non détectées, I le nombre de régions insérées, S_{cla} le nombre de régions d'erreurs correctement détectées mais mal classées, S_{reg} le nombre de régions d'erreurs dont les frontières ont été mal détectées mais assignées avec la classe d'erreur correcte, et S_{all} le nombre de régions d'erreurs dont les frontières ont été mal détectées et assignées avec une classe d'erreur incorrecte. En fonction de l'application visée, toutes les erreurs n'ont pas le même impact sur le score SER. Ici, les erreurs S_{cla} et S_{reg} ont un coût de 0,5.

Le SER obtenu pour la méthode *fusion* est de 81,6 % contre 86,7 % pour la méthode *Base*.

4.3 Impact du filtrage des mots de très faible confiance

Dans cet article, nous nous intéressons à l'impact des mots filtrés *a posteriori* par les systèmes de transcription (voir partie 2.1) sur notre méthode. Pour ce faire, nous avons utilisé tous les mots transcrits (aucun filtrage) pour entraîner les modèles des différentes approches puis avons appliqué ces méthodes sur les transcriptions de test toujours sans aucun filtrage. Enfin, pour avoir des résultats comparables, un filtrage des mots est effectué avant analyse des résultats afin de retrouver la transcription d'origine obtenant les meilleurs taux d'erreur-mots.

Comme illustré dans le tableau 2, cette approche améliore fortement le rappel de la détection, avec un gain de 13,9 points en absolu en conservant une précision de détection et un taux de classification correcte très proches. Les deux dernières colonnes présentent les résultats globaux obtenus sur la détection et la caractérisation des régions d'erreurs liées aux noms de personnes

(*régions NP*). Là encore, le taux de rappel est bien meilleur en utilisant les transcriptions non filtrées. Le système permet ainsi de détecter et correctement caractériser 40,8 % des régions *NP*. La difficulté de cette tâche s'explique en partie par le fait que nous sommes dans un problème à données déséquilibrées : seulement 5 % des régions d'erreurs sont dues à cette classes particulière. Dans le cas où toutes les régions d'erreurs détectées étaient étiquetées en *NP*, le rappel atteindrait 61,7 % mais avec une précision de 2,7 %.

TABLE 2 – Impact de l'utilisation des mots filtrés avec la méthode *Fusion*

Corpus JT test	**Détection**		**Caractérisation**	**Global**	**Régions NP**	
	Rappel	**Précision**	**% classif correcte**	**SER**	**Rappel**	**Précision**
JT train	42,2	57,0	78,4	81,6	30,4	33,3
JT train non filtré	56,1	55,0	77,2	81,0	40,8	32,2

5 Influence de la base d'apprentissage

La deuxième partie de nos expériences s'est concentrée sur l'évaluation de la détection et caractérisation des régions d'erreurs dans le cadre de données hétérogènes. En effet, nous souhaitons connaître l'impact sur les performances de cette tâche particulière lors de l'utilisation de données d'apprentissage différentes des données de test. En nous appuyant sur les résultats obtenus dans la partie 4.3, nous avons choisi d'utiliser des transcriptions non filtrées ainsi que la méthode *Fusion* et de nous focaliser particulièrement sur la classe d'erreurs *Nom de personne (NP)*. Nous évaluons la performance de cinq bases d'apprentissage sur les données *JT test* et *REPERE test* : *REPERE train*, *JT train*, *JT train reduc*, *JT train+REPERE train* et *JT train+JT test+REPERE train* (plus de détails sur ces données dans la partie 2.1). La base *JT train reduc* est une version réduite de *JT train* de taille comparable à *REPERE train*. Afin de palier au manque de données du corpus *REPERE*, l'évaluation se fait en *leave-one-out* lorsque nous utilisons *REPERE train*.

TABLE 3 – Performances sur la classe *NP* en fonction de la base d'apprentissage

Régions NP	**JT test**		**REPERE test**	
	Rappel	**Précision**	**Rappel**	**Précision**
REPERE train	17,5	19,4	24,9	28,8
JT train reduc	31,7	26,2	15,2	23,7
JT train	40,8	32,2	14,3	21,7
JT train+REPERE train	43,3	32,6	23,2	19,7
JT train+JT test+REPERE train			25,7	30,1

Pour le corpus *JT test*, nous constatons dans le tableau 3 qu'à taille de données équivalente, *JT train reduc* permet d'obtenir des résultats bien supérieurs à ceux obtenus avec *REPERE train*. Cette différence est liée à la proximité des données d'apprentissage et de test, les données *JT train* étant très proches de *JT test*. L'utilisation de *JT train* en entier permet d'améliorer encore les performances. Enfin, le corpus combiné de *JT train+REPERE train* conduit a une légère amélioration, en permettant un rappel global de 43,3 % et une précision de 32,6 % sur la détection des régions d'erreurs de noms de personnes.

Des conclusions relativement similaires peuvent être tirées sur le corpus *REPERE test*. Des données d'apprentissage proches des données de test permettent d'obtenir de meilleures performances de détection des régions d'erreurs pour le cas des *NP* (*REPERE train*) en comparaison à un corpus d'apprentissage plus éloigné (*JT train*). Ces résultats confirment l'intérêt d'une méthode à base d'apprentissage exploitant le contexte d'apparition des erreurs. L'apprentissage au moyen des

corpus *JT train+JT test+REPERE train* permet au final un léger gain pour atteindre 25,7 % en rappel et 30,1 % en précision. À données d'apprentissage équivalentes, les résultats sur ce corpus sont inférieures à celles obtenues sur le corpus *JT test*. La plus grande hétérogénéité des émissions du corpus *REPERE* explique des performances en retrait. Notons finalement que si l'on s'intéresse aux résultats globaux sur les 4 classes considérées, les évolutions des performances suivent la même tendance que celles observées sur la classe des noms de personnes.

6 Conclusion et perspectives

Dans cet article, nous nous sommes intéressés à la détection et caractérisation de régions d'erreurs dans des transcriptions automatiques de contenus multimédia. Nous considérons ce double problème comme une tâche d'étiquetage de séquences. Différentes approches ont été proposées, avec des approches dites *séquentielles* où les régions d'erreurs sont dans un premier temps détectées pour ensuite être caractérisées en classes d'erreurs, et une approche *intégrée* où ces deux problèmes sont traités conjointement. Parmi les 4 classes d'erreurs que nous cherchons à détecter, les erreurs dues à des noms de personnes ont été particulièrement étudiées car cette classe est essentielle dans de nombreuses applications d'extraction d'information. Nous avons proposé, dans un premier temps, d'étudier l'impact des mots à très faibles mesures de confiance sur notre problème d'étiquetage de séquences. Bien que ces mots soient généralement retirés de la transcription finale, leur prise en compte dans la modélisation du problème améliore les performances, particulièrement pour le rappel des régions d'erreurs détectées. Dans la seconde partie de nos expériences, nous avons cherché à comparer les performances de notre système sur deux corpus différents en faisant varier la taille et la nature des données d'apprentissage. Les résultats obtenus ont confirmé l'intérêt d'une méthode à base d'apprentissage exploitant le contexte d'apparition des erreurs. Nos travaux futurs s'orienteront vers l'utilisation de ces régions d'erreurs détectées afin de réaliser des traitements spécifiques dans la problématique d'indexation de documents, en proposant notamment des stratégies de correction.

Références

Dufour, R., Damnati, G. et Charlet, D. (2012). Automatic error region detection and characterization in lvcsr transcriptions of tv news shows. *In ICASSP*, Kyoto, Japon.

Duta, N., Schwartz, R. et Makhoul, J. (2006). Analysis of the Errors Produced by the 2004 BBN Speech Recognition System in the DARPA EARS Evaluations. *In IEEE TASLP*, volume 14, pages 1745–1753.

Gauvain, J.-L., Lamel, L. et Adda, G. (2002). The LIMSI Broadcast News Transcription System. *Speech Communication*, pages 89–108.

Lafferty, J., McCallum, A. et Pereira, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In ICML*, Williamstone, États-Unis.

Makhoul, J., Kubala, F., Schwartz, R. et Weischedel, R. (1999). Performance measures for information extraction. *In Darpa broadcast news workshop*.

Parada, C., Dredze, M., Filimonov, D. et Jelinek, F. (2010). Contextual information improves OOV detection in speech. *In NAACL-HLT*, Los Angeles, États-Unis.

Rastrow, A., Sethy, A. et Ramabhadran, B. (2009). A new method for OOV detection using hybrid word/fragment system. *In ICASSP*, pages 3953–3956, Taipei, Taiwan.

Vasilescu, I., Adda-Decker, M., Lamel, L. et Halle, P. (2009). A perceptual investigation of speech transcription errors involving frequent near-homophones in French and American English. *In Interspeech*.

Woodland, P., Johnson, S., Jourlin, P. et Spärck Jones, K. (2000). Effects of out of vocabulary words in spoken document retrieval. *In SIGIR*, pages 372–374, Athènes, Grèce.