Corrections spécifiques du français sur les systèmes de reconnaissance automatique de la parole

Richard Dufour, Yannick Estève, Paul Deléglise

LIUM, Université du Maine
avenue René Laënnec - 72083 Le Mans, France
Tél. : 02 43 83 38 43 - Fax : 02 43 83 38 68
Courriel : prénom.nom@lium.univ-lemans.fr

ABSTRACT

Automatic speech recognition (ASR) systems are used in a large number of applications, in spite of the inevitable recognition errors. In this study we propose a pragmatic approach to automatically repair ASR outputs by taking into account linguistic and acoustic information, using formal rules or stochastic methods. The proposed strategy consists in developing a specific correction solution for each specific kind of errors. In this paper, we apply this strategy on two case studies specific to French language. We show that it is possible, on automatic transcriptions of French broadcast news, to decrease the error rate of a specific error by 11.4% in one of two the case studies, and 86.4% in the other one. These results are encouraging and show the interest of developing more specific solutions to cover a wider set of errors in a future work.

Index Terms : Reconnaissance automatique de la parole, correction d'erreurs, homophones, méthode statistique

1. INTRODUCTION [1]

Les systèmes de reconnaissance automatique de la parole (RAP) sont de plus en plus efficaces. Leurs performances actuelles sont suffisantes pour qu'ils soient utilisés dans de nombreuses applications (Dialogue Homme-Machine, indexation, recherche d'informations...).

Mais les erreurs de reconnaissance sont inévitables. Les erreurs modifiant le sens de la phrase sont pénibles car elles empêchent des retours corrects de la part des utilisateurs. En revanche, d'autres erreurs ne gênent pas la compréhension et sont souvent négligées car elles ne sont pas critiques pour réaliser une opération (nous pensons notamment aux erreurs d'accord).

Pour certaines applications, comme le sous-titrage ou les transcriptions assistées [Baz08], ces erreurs sont plus importantes : leurs répétitions, même si elles ne modifient pas le sens, est très fatiguant pour l'utilisateur.

La langue française contient de nombreux mots homophones, notamment à travers les formes fléchies d'un même mot. Les erreurs fréquentes de RAP en résultent. Dans ce contexte, il serait intéressant d'avoir une méthode pour les corriger.

Dans la littérature, nous trouvons des propositions pour réparer les erreurs des systèmes de RAP ou rendre les applications robustes à ces méthodes [Wal00, Ser06]. En général, ces propositions cherchent à réparer de manière générale toutes les sortes d'erreurs [Sag04]. Des propositions prennent en compte des particularités de l'application utilisée, avec par exemple l'historique du dialogue [Sag04].

Dans ce papier, nous proposons une approche différente : celle-ci a pour but de construire une correction spécifique pour chaque type d'erreurs. Une idée similaire a été proposée dans [Est02] pour une utilisation à l'intérieur du système de reconnaissance automatique de la parole en traitant différents modèles de langage. Nous proposons de l'utiliser pour réparer les erreurs en "post-traitant" les sorties textes des systèmes de RAP.

L'approche consiste à analyser de manière manuelle les erreurs les plus fréquentes, particulièrement les paires de confusion. Les erreurs peuvent être vues comme un groupe, ou être traitées de manière isolée. Les différentes solutions proposées pour les différentes sortes d'erreurs peuvent être soit des règles formelles, soit des méthodes stochastiques. Ces outils peuvent venir de données variées (connaissances acoustiques ou linguistiques). L'utilisation de l'information acoustique pour corriger les sorties de RAP en post-traitement est une autre contribution de nos travaux.

La section suivante présente quelques particularités du français. Puis l'approche proposée sera détaillée dans la section 3, avant la présentation des outils utilisés. Enfin, les expériences seront décrites et les résultats seront présentés.

2. QUELQUES PARTICULARITÉS DU FRANÇAIS

La flexion en genre et en nombre est l'un des aspects les plus difficiles du français. Une grande difficulté des systèmes de RAP est que les différentes formes fléchies d'un mot sont souvent homophones et seul le modèle de langage du RAP peut les distinguer.

L'accord en genre et en nombre n'est pas toujours bien modélisé par les RAP car, d'une part la longueur des contraintes modélisées par les modèles n-gram peut ne pas être suffisante, et d'autre part à cause de la faiblesse du modèle de langage et du manque de données : [Gau94] montre que pour obtenir la même couverture de mots en français qu'en anglais au niveau du vocabulaire d'un système de RAP, il est nécessaire d'avoir deux fois plus de mots en français.

[1] Recherche financée par l'*Agence Nationale de la Recherche* sous contrat ANR-06-MDCA-006.

De plus, de nombreuses règles grammaticales complexes ne peuvent être modélisées avec un modèle de langage n-gram.

3. APPROCHE PROPOSÉE

Dans ce papier, nous nous intéressons aux erreurs causées par les formes fléchies homophones des participes passés, ainsi qu'aux erreurs des mots *'vingt/vingts'* et *'cent/cents'*. Ces erreurs sont quelques unes des plus fréquentes produites par les RAP, après analyse avec l'outil **NIST SCLITE**.

Pour corriger les erreurs, nous cherchons à appliquer des règles formelles. Cette méthode consiste à utiliser une règle linguistique stricte et à la modéliser de manière à pouvoir l'appliquer directement sur les sorties d'un décodeur. Afin de pouvoir mettre en place cette méthode, il faut être certain que la règle linguistique est assez simple et précise pour fonctionner dans la grande majorité des cas. Ainsi, par exemple, les erreurs d'accord sur les mots *'cent'* et *'vingt'* sont potentiellement corrigibles au moyen d'une règle formelle : en effet, les règles concernant ces accords possèdent un cadre bien défini, et la mise en place informatique est simple.

Mais nous constatons que cette approche n'est pas suffisante. En particulier, les règles formelles ne sont pas très robustes aux erreurs se trouvant dans le contexte lexical d'un mot. Ainsi, lorsque cela est possible, nous utilisons une règle formelle, sinon nous utilisons une méthode statistique, par exemple un classifieur construit à partir d'un corpus d'apprentissage, pour remettre en cause la lexie du mot fourni par le système de RAP et, le cas échéant, le corriger.

Nous ne cherchons pas à corriger toutes les erreurs, mais seulement les mots qui semblent corrigibles. Pour développer une solution spécifique, plusieurs bases de connaissances peuvent être utilisées : information lexicale, mais aussi acoustique, texte étiqueté, ou d'autres niveaux d'information.

Le texte étiqueté est très utile pour construire des règles formelles et pour donner au classifieur des informations plus riches. L'information acoustique peut être la prononciation utilisée par le système de RAP pour reconnaître un mot. Il est possible de lier la prononciation à d'autres informations linguistiques (genre et nombre) pour guider les choix de correction.

Prenons l'exemple d'une correction proposée par l'outil statistique sur le mot "transcrit" (masculin/singulier) ayant les phonèmes [t r an s c r i]. En admettant que le mot est considéré faux par notre classifieur et qu'il propose la forme féminin/singulier. Cela donnerait "transcrite" [t r an s c r i **t**]. Mais l'information acoustique est différente de la proposition du système de RAP, la proposition n'est donc pas retenue. Ainsi, l'information acoustique permet parfois de trouver le mot discriminant de par sa forme acoustique. En effet, une forme fléchie d'un mot peut contenir l'information sur le genre et le nombre, et permet alors de corriger le mot erroné. Selon les informations fournies au classifieur durant la phase d'apprentissage, celui-ci pourra sélectionner la classe la plus adaptée.

Les classes sont déterminées en fonction des mots ciblés à réparer. Par exemple, si nous voulons connaître le nombre d'un mot, le singulier sera considéré comme une classe et le pluriel comme une autre. Le classifieur testera le mot et lui proposera un nombre. Quand le classifieur et le système de RAP ne sont pas d'accord, nous remplaçons le mot "hypothèse" du RAP par la forme fléchie correspondant au nombre donné par le classifieur.

L'approche est pragmatique et divisée en deux parties, d'une part la mise en place de règles formelles, et d'autre part l'utilisation de méthodes statistiques (classification) sur des problèmes spécifiques. La méthode donnant les meilleurs résultats est alors conservée.

4. OUTILS

Les expériences sur la reconnaissance de la parole ont été effectuées sur le système de RAP du LIUM, basé sur le décodeur CMU Sphinx 3.x, décrit dans [Del05]. Ce décodeur est un système en trois passes : la première passe utilise un modèle de langage trigramme et des modèles acoustiques génériques (un pour chacune des quatre conditions selon le genre/bande — homme/femme + studio/téléphone) ; la seconde passe utilise la meilleure hypothèse de la première passe pour adapter les modèles acoustiques en utilisant SAT et CMLLR ; la dernière passe consiste à réévaluer un graphe de mots généré pendant la seconde passe au moyen d'un modèle de langage quadrigramme. Le système a été classé deuxième à la campagne d'évaluation ESTER [Gal05] sur les RAP français, et était le meilleur système open source. Sur les données de test, le système a un taux d'erreurs de 22,2%. Le taux d'erreurs au niveau des substitutions est de 13,6% : ce taux est intéressant pour nos expériences car l'approche proposée dans ce papier a pour vocation de réparer seulement les erreurs de substitution (les insertions et suppressions ne sont pas ciblées ici).

Le texte étiqueté est associé avec les mots hypothèses en utilisant l'outil de tagging **lia_tagg**[2], distribué sous licence GPL.

Nous utilisons le classifieur à l'état de l'art **BoosTexter** [ES00] pour notre méthode statistique. Ce classifieur implémente l'algorithme **AdaBoost**, qui permet de construire une puissante classification à partir de chaînes de caractères et de valeurs continues.

5. DONNÉES EXPÉRIMENTALES

Les données utilisées pour les expériences sont issues de la campagne d'évaluation ESTER sur la transcription de journaux radiophoniques français, correspondant à 100 heures d'enregistrements audio transcrits.

Le système de reconnaissance automatique de la parole utilisé pour les expériences a été développé en prenant les données transcrites provenant de la campagne ESTER.

Les données sont divisées en trois corpus : données d'*apprentissage*, données de *développement* et données de *test* (le tableau 1 montre les tailles respectives des corpus).

Le classifieur statistique a été construit à partir des transcriptions manuelles des données d'apprentissage. Notons

[2]http ://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/bechet/download_fred.html

que les modèles de langage utilisés dans le système de RAP étaient, en petite partie, estimés à partir des mêmes données d'apprentissage (et avec des articles de journaux issus du journal "Le Monde", non utilisés pour entraîner le classifieur).

Table 1: Taille (mots et temps) des trois corpus.

	Apprentissage	*Développement*	*Test*
Words	840K	87K	114K
Time	73 heures	7 heures	10 heures

L'information acoustique (phonèmes associés aux mots) sont dans le dictionnaire de prononciation du décodeur. Le dictionnaire contient environ 165K variantes de prononciations, pour un total de 62K mots.

Afin de pouvoir corriger les mots fournis par le système de RAP, toutes les formes grammaticales d'un mot sont attendues. Les informations ont été récupérées sur la base de données **lexique3**[3], contenant 135K mots. La base donne la représentation en genre, nombre, syllabes et lemmes associés.

6. EXPÉRIENCES

Nous avons choisis les erreurs fréquentes des systèmes de reconnaissance automatique de la parole. Ainsi, en premier lieu, nous avons cherché à modéliser la même règle formelle sur les mots *cent* (singulier *cent* / pluriel *cents*) and *vingt* (singulier *vingt* / pluriel *vingts*) car le système de reconnaissance automatique de la parole modélise très mal cette règle d'accord.

Dans une seconde étape, nous nous sommes intéressés aux erreurs d'accord sur les participes passés, qui peuvent prendre jusqu'à quatre formes concurrentes selon le genre et le nombre. La difficulté réside dans le fait que ces formes peuvent le plus souvent être phonétiquement identiques. Il apparaît donc extrêmement difficile d'appliquer une règle grammaticale précise pour les corriger. Nous avons d'abord étiqueté (au moyen du tagger présenté dans 4) les mots en contexte gauche du participe passé. Ces informations ont été fournies pendant la phase d'apprentissage afin de généraliser les événements observés. Cet apprentissage permettra au classifieur de fournir le genre et le nombre d'un mot en fonction des données de test choisies.

Si le genre et le nombre fournis par le classifieur sont différents de ceux proposés par le système de RAP, la proposition du classifieur est retenue si, seulement si, la prononciation de la nouvelle forme fléchie est équivalente. Ainsi, nous ne remettons pas en cause les mots et les phonèmes associés par le système de RAP.

7. RÉSULTATS

Afin de savoir si notre approche aide à corriger les erreurs ciblées, nous avons comparé les taux d'erreurs sur les sorties de notre système de RAP avec les taux d'erreurs sur les sorties corrigées au moyen de notre méthode. Le premier taux d'erreurs a été effectué sur le corpus de **développement**, afin de calibrer notre système, et le second sur le corpus de **test**, pour s'assurer de son efficacité. Sur le tableau 2 nous présentons les taux d'erreurs d'accord des mots "cent" et "vingt" avant (Baseline) et après correction au moyen de la règle linguistique (Correction).

Table 2: Taux d'erreurs d'accord sur les mots "cent" et "vingt".

	Développement	*Test*
Baseline	16,7%	6,6%
Correction	0,6%	0,9%
Gain relatif	**96,7%**	**86,4%**

Nous constatons que l'impact de cette règle formelle est positif. Pendant la phase de développement, le gain était très haut, avec une correction de 87 mots sur un total de 90. Nous avons alors testé notre système sur les données de **test** et avons observé que seulement 6 mots étaient toujours mal reconnus après correction (dû à un manque de données au niveau de l'historique ou à un choix erroné du mot par le système de RAP). Nous avons alors réussi à corriger 28 erreurs sur la forme "cent/cents" et 13 sur la forme "vingt/vingts". Bien que les gains soient moins importants pendant la phase de **test** que pendant la phase de **développement**, ils restent néanmoins intéressants : en utilisant une règle linguistique en "post-traitement" des sorties de système de reconnaissance automatique de la parole, nous pouvons constater qu'il est possible de corriger des erreurs d'accord, avec un gain relatif de 86,4% sur les erreurs de ces mots.

Intéressons-nous maintenant aux résultats de notre seconde expérience : la correction de formes fléchies homophones sur les participes passés en utilisant une méthode statistique. Le tableau 3 compare le taux de participes passés ayant une erreur d'accord avant et après l'utilisation de la méthode statistique.

Table 3: Taux d'erreurs d'accord dus aux participes passés.

	Développement	*Test*
Baseline	10,1%	12,9%
Correction	8,1%	11,4%
Gain relatif	**19,6%**	**11,4%**

Etant donné les résultats positifs que nous avons eu sur les données de *développement* (58 mots corrigés sur un total de 296), nous avons décidé d'étendre comme précédemment notre système sur les données de *test*. Notons que les meilleurs résultats ont été constatés en entraînant le classifieur durant 800 tours, et en utilisant l'option *n*-gram avec une longueur de 4. Ainsi, sur les données de *test*, le classifieur a détecté 253 participes passés contenant des erreurs d'accord. Grâce à l'utilisation de paramètres acoustiques, 80 participes passés (31,62%) devant être modifiés ont été ignorés. Cette contrainte acoustique a permis d'améliorer le gain relatif de 2,6 points.

La figure 1 résume les performances de la méthode statistique couplée à des contraintes acoustiques sur les parti-

[3]http ://www.lexique.org

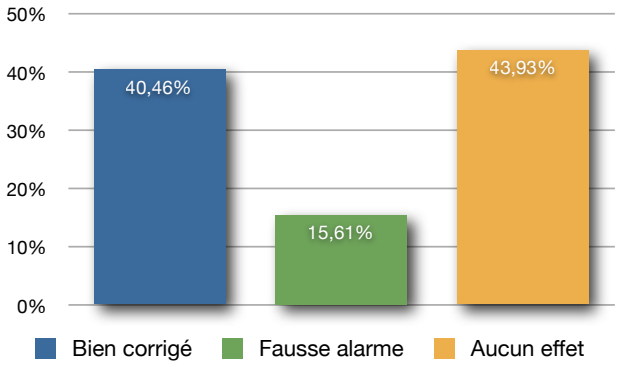

Figure 1: Proportion d'erreurs d'accord due aux participes passés correctement et mal modifiés avec la méthode statistique.

cipes passés modifiés. Nous pouvons voir les erreurs bien corrigées, les fausses alarmes (mots bien décodés mais mal corrigés) et les "aucun effet" (mots mal décodés et mal corrigés).

Nous notons que la proportion des participes passés correctement modifiés (70 mots) est bien plus grande que ceux mal corrigés (avec insertion d'erreurs) de participes passés (27 mots). Au final, la méthode a permis de corriger 43 participes passés affectés d'une erreur d'accord (gain relatif de 11,41%). La méthode modifie également la flexion de participes passés erronés, sans réussir à bien les corriger : aucun impact n'est cependant visible puisque l'erreur du décodeur persiste après l'utilisation de la méthode (43,93% des mots corrigés).

Ces méthodes (règles linguistiques, et méthodes statistiques combinées aux paramètres acoustiques) permettent de corriger des erreurs d'accord sur les transcriptions fournies par un système de RAP sans avoir besoin de modifier le décodeur. Ce travail préliminaire, ciblé sur un nombre limité de règles, montre qu'il est possible de corriger ces types d'erreurs. Bien sûr, ces travaux doivent être élargis, mais un gain dans ce type de correction semble réalisable.

De plus, un autre avantage de ces méthodes est le temps de calcul relativement faible. En effet, appliquer ces méthodes ne prend qu'environ 6 minutes, ce qui est négligeable par rapport à la quantité de données traitées.

8. Conclusion

Dans ces travaux, nous avons proposé une stratégie consistant à créer des solutions spécifiques pour palier les erreurs des systèmes de reconnaissance automatique de la parole. Nous avons appliqué cette stratégie à deux cas d'étude qui figurent parmi les erreurs les plus fréquentes de ces systèmes.

Nous avons tout d'abord proposé une solution de correction d'erreurs au moyen d'une règle linguistique sur les mots homophones *'vingt/vingts'* et *'cent/cents'*. Cette solution utilise les règles formelles et permettent de réduire le taux d'erreurs sur ces mots de 86,4% sur le corpus ESTER.

Puis, nous avons proposé une solution pour corriger les erreurs dues aux formes fléchies des participes passés. Dans ce cas, la méthode stochastique utilisant le classifieur *BoosTexter* permet de réduire le taux d'erreurs sur ce type d'erreurs de près de 11%. Nous essayons au moyen cette méthode statistique de corriger des erreurs d'accord (très souvent des homophones) dues modèle de langage, en réalisant une correction des sorties de systèmes de RAP. L'idée de cette méthode est de notamment utiliser des informations non présentes dans le modèle de langage, à savoir l'étiquetage des rôles grammaticaux.

Bien sûr, ces deux améliorations ne sont pas suffisantes pour réduire de manière significative le taux d'erreurs global. Mais nous avons montré que nous pouvons réduire de manière drastique le taux d'erreurs sur ce type d'erreurs identifiées. Nous avons constaté, en analysant les erreurs les plus fréquentes de notre système de RAP, qu'il était possible de réduire le taux d'erreurs de substitution d'environ 21,5% en relatif en ne prenant en compte que les 30 erreurs de substitution les plus fréquentes susceptibles d'être corrigées par notre système.

Dans de futurs travaux, nous nous intéresserons à d'autres erreurs spécifiques, afin d'agrandir notre champs de correction. Nous nous efforcerons également à améliorer notre système, notamment sur la méthode statistique, en prenant en compte un plus grand nombre d'informations au niveau de l'historique, et en utilisant de manière plus avancée les informations acoustiques fournies par le système de RAP.

Enfin, les méthodes de post-traitement mises en place ont été conçues de manière à être indépendantes du système de RAP utilisé. Nous testerons la réutilisabilité de notre approche sur d'autres systèmes.

Références

[Baz08] Bazillon, T., Estève, Y. and Luzzati, D. (2008), "Manual vs assisted transcription of prepared and spontaneous speech", in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

[Del05] Deléglise, P., Estève, Y., Meignier, S. and Merlin, T. (2005), "The LIUM speech transcription system : a CMU Sphinx III-based System for French Broadcast News", in *Interspeech*, Lisbon, Portugal.

[ES00] E. Schapire, R. and Singer, Y. (2000), "Boostexter : A boosting-based system for text categorization", *Machine Learning*, vol. 39, pp. 135–168.

[Est02] Estève, Y., Raymond, C. and De Mori, R. (2002), "On the use of structures in language models for dialogue : Specific solutions for specific problems", in *ISCA TRW on Multi-modal dialogue in mobile environments*, Kloster Irsee, Germany.

[Gal05] Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J. and Gravier, G. (2005), "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news", in *Interspeech*, Lisbon, Portugal.

[Gau94] Gauvain, J.L., Lamel, L., Adda, G. and Adda-Decker, M. (1994), "Speaker-independent continuous speech dictation", *Speech Communication*, vol. 15.

[Sag04] Sagawa, H., Mitamura, T. and Nyberg, E. (2004), “Correction grammars for error handling in a speech dialog system”, in *Human Language Technology conference (HLT/NAACL 04)*, Boston, MA, USA.

[Ser06] Servan, C., Raymond, C., Béchet, F. and Nocéra, P. (2006), “Conceptual decoding from word lattices : application to the spoken dialogue corpus media”, in *Proceedings of the International Conference on Spoken Language Processing (Interspeech'06)*, Pittsburgh, PA, USA.

[Wal00] Walker, M. and Langkilde, I. (2000), “Using natural language processing and discourse features to identify understanding errors in a spoken dialogue system”, in *In Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 1111–1118.