Prediction of transcription indexability

Gregory Senay, Benjamin Lecouteux, Georges Linares

Prédiction de l'indexabilité d'une transcription

Grégory Senay[1] Benjamin Lecouteux[2] Georges Linarès[1]
(1) LIA, AVIGNON (2) LIG, GRENOBLE
gregory.senay@univ-avignon.fr, benjamin.lecouteux@imag.fr,
georges.linares@univ-avignon.fr

RÉSUMÉ
Cet article présente une mesure de confiance sémantique permettant de prédire la qualité d'une transcription automatique dédiée à de la recherche d'information dans les documents audio (RIDA). La méthode proposée est basée sur une combinaison de la mesure de confiance issue du système automatique de reconnaissance de la parole (SRAP) et d'un index de compacité sémantique (ICS). Elle permet d'estimer la pertinence des mots en fonction du contexte sémantique dans lequel ils apparaissent. Les expériences sont menées sur le corpus de la campagne ESTER 2, en simulant un scénario classique d'utilisation d'un système de RIDA : les utilisateurs soumettent des requêtes textuelles à un moteur de recherche qui est supposé leur retourner les documents audio les plus pertinents. Les résultats démontrent l'intérêt d'utiliser un niveau d'information sémantique pour prédire l'*indexabilité* de la transcription.

ABSTRACT
Prediction of transcription indexability

This paper presents a semantic confidence measure that aims to predict the relevance of automatic transcripts for a task of Spoken Document Retrieval (SDR). The proposed predicting method relies on the combination of Automatic Speech Recognition confidence measure and a Semantic Compacity Index, that estimates the relevance of the words considering the semantic context in which they occurred. Experiments are conducted on the French Broadcast news corpus ESTER 2, by simulating a classical SDR usage scenario : users submit text-queries to a search engine that is expected to return the most relevant documents regarding the query. Results demonstrate the interest of using semantic level information to predict the transcription *indexability*.

MOTS-CLÉS : Reconnaissance de la parole, mesure de confiance, recherche d'information, document audio.

KEYWORDS: Speech recognition, confidence measures, spoken document retrieval.

1 Introduction

Les approches habituelles en recherche d'information dans les documents audio (RIDA) associent un système de reconnaissance automatique de la parole (SRAP) et des techniques de recherche d'information (RI). Un des enjeux majeurs de cette approche est l'impact des erreurs de reconnaissance sur les performances du système de RI : les SRAP ne sont pas assez robustes dans des conditions inattendues où le taux d'erreur mot (TEM) peut être supérieur à 30 % et perturber ainsi significativement la précision de la recherche (Oard *et al.*, 2004; Whittaker *et al.*, 2002;

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 697–705,
Grenoble, 4 au 8 juin 2012.

Hansen *et al.*, 2005). Dans des conditions contrôlées, la campagne TREC-SDR conclut que ces erreurs ne corrompent pas les résultats du moteur de recherche (Garofolo *et al.*, 2000).

Considérant qu'un SRAP parfait n'existera pas à court terme, plusieurs études récentes en RIDA se focalisent sur des méthodes tolérantes aux erreurs des SRAP. Elles se basent sur les différentes représentations des transcriptions (treillis de mots, N-meilleures hypothèses...) (Saraclar, 2004; l. Chang *et al.*, 2008), les stratégies d'indexation (Chelba *et al.*, 2007; Kurimo et Turunen, 2005; Siegler, 1999) ou le traitement des mots hors vocabulaire.

Pour des applications industrielles, une méthode réaliste serait d'identifier les segments de la transcription où le SRAP échoue, pas seulement en terme de TEM mais aussi en considérant l'objectif final : la recherche d'information. Ensuite ce segment erroné pourrait être vérifié et corrigé par un humain. Dans un scénario semi-automatique, la disponibilité d'un outil d'auto-diagnostique (qui peut aider à identifier les segments erronés) est critique pour le coût global du processus d'indexation. Ce papier présente une telle méthode qui a pour but de prédire à quel point une erreur de transcription peut dégrader les performances globales du système de RIDA.

Cet article est la suite de notre article (Senay *et al.*, 2011), validant les résultats obtenus sur un corpus plus récent et plus important. La section 2 décrit la méthode et la métrique d'évaluation de la qualité d'indexation d'un segment. La section 3 introduit la méthode pour prédire l'indexabilité. Le protocole expérimental est présenté dans la section 4. Le dernier chapitre présente les conclusions et les perspectives.

2 Indexabilité d'un document

L'évaluation de l'impact du TEM dans la RIDA a été abordé et étudié dans de nombreux articles (Chelba *et al.*, 2008). Généralement dans les campagnes en RIDA, les résultats générés par le système de RIDA sont comparés à un classement de référence établi par des experts. Une autre méthode consiste à comparer les classements obtenus à partir des transcriptions issues du SRAP et celles qui ont été transcrites manuellement. Ces évaluations sont effectuées en utilisant un large jeu de requêtes, soumis au moteur de recherche opérant sur l'ensemble d'un corpus de test. Les performances du système de RIDA sont calculées avec les mesures MAP (Mean Average Precision) ou R-Precision (précision au rang N).

Dans cet article, notre but est de prédire, au niveau du segment de transcription, quel est l'impact des erreurs pour le processus global de RIDA. La section suivante présente comment cette mesure de l'indexabilité est estimée.

2.1 Estimation de l'indexabilité

Les segments de transcription sont découpés automatiquement par rapport aux silences avec une durée maximum de 30 secondes. Chacun d'eux est considéré comme un document par le système de RIDA. Considérant qu'une seule erreur dans le segment peut potentiellement modifier tous les résultats de recherche (pour l'ensemble des questions), l'estimation de l'indexabilité d'un segment nécessite une évaluation individuelle.

Pour cela, l'indexabilité $Idx(s)$ d'un segment s est calculée en 3 étapes :

1. le segment ciblé s est automatiquement transcrit par le SRAP,
2. pour chacune des requêtes, une recherche est effectuée sur le corpus de référence, excepté pour s qui a été automatiquement transcrit,
3. les classements obtenus sont comparés avec ceux obtenus sur le corpus de référence. Finalement, l'indexabilité $Idx(s)$ du segment s est obtenue en calculant la F-mesure sur les 20 meilleurs résultats des classements.

Cet algorithme estime l'impact individuel du segment de transcription ciblé dans le processus global de RIDA, en connaissant *a priori* le classement du segment. La prochaine section présente une méthode pour prédire cette mesure d'*indexabilité*.

3 Prédiction de l'indexabilité

La méthode proposée aide à prédire l'impact des erreurs de reconnaissance dans le processus d'indexation. Pour cela, nous combinons des mesures de confiance au niveau du mot et un index de compacité sémantique sur la meilleure hypothèse générée par le SRAP. La combinaison est effectuée en utilisant un perceptron multi-couches. Ces principaux éléments sont décrits dans les sections suivantes.

3.1 Mesure de confiance du SRAP

Le score de confiance permet d'estimer la probabilité qu'un mot soit juste ou faux. Ce score qui est issu du SRAP est calculé dans nos expériences en 2 étapes.

La première extrait des paramètres de bas niveau relatifs à l'acoustique et au graphe de recherche du décodeur, puis des paramètres de haut niveau relatifs à la linguistique. Chaque mot de l'hypothèse est ainsi représenté par un vecteur de 23 paramètres, qui sont regroupés en 3 classes :

– **Les paramètres acoustiques** se composent de la vraisemblance acoustique du mot, la vraisemblance par trame, la différence entre la vraisemblance du mot et le score de décodage du segment sans contraintes acoustiques.
– **Les paramètres linguistiques** sont basés sur des probabilités estimées par le modèle de langage utilisées dans le SRAP. Nous utilisons les probabilités avec un modèle 3-grammes, la perplexité du mot dans son contexte et la probabilité unigramme. Nous ajoutons une information renseignant sur le comportement de repli du modèle de langage.
– **Les paramètres issus du graphe de décodage** sont basés sur l'analyse du réseau de confusion : le nombre de chemins alternatifs pour un mot et les valeurs relatives à la distribution des probabilités *a posteriori*.

Dans la seconde étape, un classifieur basé sur un algorithme de *boosting* attribue une probabilité de rejet ou non du mot comme détaillé dans (Moreno *et al.*, 2001). L'algorithme consiste en une recherche exhaustive pour une combinaison linéaire en surpondérant les exemples mal classés. Le classifieur est entraîné sur un corpus d'entraiment spécifique qui n'a pas été inclus dans l'entraînement du SRAP. Chaque mot de ce corpus est étiqueté comme *correct* ou *erroné*, selon la référence du SRAP.

Cette méthode permet d'obtenir une mesure de confiance pour chacun des mots du segment. Le paramètre de prédiction de référence est calculé en faisant la moyenne des scores de confiance des mots du document porteurs de sens (filtrés à l'aide d'une stop-liste de 729 mots contenant principalement des articles, des adjectifs démonstratifs, des adjectifs possessifs...), permettant d'obtenir une mesure de confiance au niveau du segment. Cette méthode de référence est utilisée pour entrainer un perceptron à une seule entrée pour prédire l'indexabilité.

Cette mesure de confiance obtient une Entropie Croisée Normalisée (NCE) de 0,373 sur le corpus de développement et de 0,282 sur le test. Cette NCE a été calculée directement sur les transcriptions générées par le SRAP.

3.2 Index de Compacité Sémantique

L'utilisation d'une information de niveau sémantique pour la prédiction de l'indexabilité est motivée par le fait qu'une requête cible en général les documents selon leurs contenus sémantiques (sujets ou bien des concepts plus fins). Lors de sa requête, l'utilisateur veut cibler des documents selon leurs thèmes. Plusieurs articles proposent d'utiliser des paramètres de haut niveau pour estimer les mesures de confiance (Cox et Dasmahapatra, 2002; Hakkani-Tür *et al.*, 2005). La plupart des auteurs concluent que ces approches n'améliorent pas significativement ni systématiquement la précision des mesures de confiance pour les SRAP. Néanmoins, les mots pertinents sont critiques pour la recherche dans les documents audio et le TEM n'évalue pas la fidélité sémantique des transcriptions.

Notre proposition est d'estimer un score de compacité sémantique $ISC(s)$ pour chacun des segments s et d'utiliser celui-ci en tant que paramètre de prédiction. Le score du segment est obtenu en moyennant localement les corrélations sémantiques $sc(w_i, w_j)$ des paires de mots (w_i, w_j) estimées sur un large corpus.

Cette approche se base sur une corrélation à court terme entre les mots porteurs de sens, d'où un filtrage des mots outils (à l'aide de la même stop-liste précédemment utilisée). De plus, les termes sont lemmatisés que ce soit pour le corpus où s'effectue la recherche ou les segments issus du SRAP. Enfin, les scores sémantiques des paires de mots sont calculés en utilisant la fréquence des cooccurrences de lemmes pondérée par un index TF-IDF :

$$cs(w_i, w_j) = TF(l_i, c).IDF(l_i).\delta^c(w_j) + TF(l_i, c).IDF(l_j).\delta^c(w_i) \quad (1)$$

Où l_i est le lemme du mot w_i, $TF(l_i, c)$ la fréquence du lemme l_i dans le contexte c, $IDF(l_k)$ la fréquence inverse du lemme l_k sur l'ensemble du corpus et δ est une valeur booléenne définie par $\delta^c(w_i) = 1$ si $w_i \in c$, 0 sinon.

Les compacités sémantiques $ics(c)$ sont estimées avec une fenêtre glissante de k lemmes, chacun correspondant à un contexte c.

$$ics(c) = \sum_{c_k} \sum_{(w_i, w_j) \in c_k} \sqrt{cs(w_i, w_j) * \frac{IDF(w_i)IDF(w_j)}{\sum_{k=1}^{n} IDF(w_k)}} \quad (2)$$

Dans nos expériences, ces scores sont calculés sur le corpus français *Wikipédia* qui offre l'avantage de couvrir un large panel de sujets et thèmes.

3.3 Combinaison des scores

Les mesures de confiance du SRAP et l'index de compacité sémantique sont combinés pour prédire le score d'indexabilité du segment. La combinaison est effectuée avec un réseau de neurones de type perceptron multicouche (Rosenblatt, 1962) qui utilise un algorithme de rétropropagation des erreurs. Les couches d'entrée, du milieu et de sortie sont respectivement de deux, dix et une cellules.

4 Protocole expérimental

4.1 Corpus

Les expériences sont conduites sur la base de données de la campagne ESTER2. Elle est composée d'enregistrements (entre les années 2007 et 2008) de journaux d'information radiophoniques, manuellement transcrits. Le corpus de développement (6 heures - 1988 segments) est utilisé pour apprendre la prédiction de l'indexabilité. Le corpus de test est utilisé (6 heures - 2362 segments) pour valider la phase d'apprentissage.

4.2 Système de reconnaissance

Les expériences utilisent le moteur de reconnaissance de la parole du LIA *SPEERAL*. Le lexique contient environ 85000 mots. Le processus complet s'effectue en 3 passes incluant une adaptation en locuteur non supervisée et un post décodage des réseaux de confusion avec un modèle de langage en 4-grammes. Nous utilisons dans nos expériences, les résultats du système obtenus lors de la seconde passe. Le système obtient un taux d'erreur mot de 26,84% sur les 6 heures du test d'ESTER 2.

4.3 Moteur de recherche et jeu de requêtes

Notre but étant d'évaluer la qualité des données plutôt que la stratégie de recherche, nous utilisons un moteur de recherche (basé TF-IDF) fréquemment utilisé *Lucene* (Hatcher et Gospodnetic, 2004). Il nous permet de retourner la liste ordonnée des documents qui sera utilisée pour calculer l'indexabilité de ces derniers. Le jeu de requêtes est construit à partir des titres du journal *Le Monde* et de plusieurs almanachs disponibles sur internet des années 2007 et 2008 *(RFI : rétrospective 2007 et 2008 ; Le Figaro : ils ont marqué l'année 2007 et 2008 ; Wikipédia : les événements de 2007 et 2008)*. En tout, le jeu de requêtes est composé de 63000 requêtes uniques, chacune d'elle produisant au moins un résultat de recherche.

5 Expériences

La première expérience a pour but d'évaluer l'erreur de prédiction de l'indexabilité (PER). Au lieu d'estimer l'impact individuel de chacun des paramètres, nous entraînons le réseau de neurones basé sur les mesures de confiance (MC), l'index de compacité sémantique (ICS) et la combinaison des deux (MC et ICS), noté $MC + ICS$.

La distorsion PER entre l'indexabilité Idx et la prédiction de l'indexabilité $Pidx$ est évaluée selon deux mesures :

$$D = \frac{1}{\tau} \sum_{j=1}^{\tau} \frac{|PIdx(j) - Idx(j)|}{Idx(j)} \tag{3}$$

$$RMS = \sqrt{\frac{1}{\tau} \sum_{j=1}^{\tau} (PIdx(j) - Idx(j))^2} \tag{4}$$

D et RMS représentent respectivement la distorsion générale et la déviation standard (RMS - Root Mean Square).

Dans nos résultats, le score $MC + ICS$ est significativement plus performant que les deux métriques individuelles. Nous pouvons voir que l'écart absolu avec la prédiction sémantique est meilleur qu'avec la mesure de confiance (17% relatif).

	MC	ICS	$MC + ICS$
D	16,88	15,11	14,38
RMS	21,46	20,85	20,25

TABLE 1 – Ce tableau présente les résultats obtenus en erreur de prédiction de l'indexabilité en utilisant respectivement la mesure de confiance (MC), la mesure de compacité sémantique (ICS) et la combinaison des deux ($MC + ICS$).

Dans la seconde expérience, nous vérifions l'intérêt des méthodes proposées pour la prédiction de l'indexabilité d'un document dans un scénario particulier où la métrique est supposée indiquer, à un archiviste, les segments qui pourraient être manuellement corrigés (afin de les rendre correctement indexables). C'est une tâche de classification de documents où chaque document est étiqueté comme indexable ou non indexable par le système.

Nous estimons les performances de la classification en comparant les deux classes. Nous utilisons le score d'indexabilité de référence et celui qui est prédit selon un seuil T. Un document est étiqueté au final comme bien classifié, seulement si son indexabilité et la prédiction de son indexabilité sont tous les deux inférieurs ou tous les deux supérieurs au même seuil T. Ce seuil varie entre 10% et 90%. Le score de classification est estimé classiquement avec une $F - mesure$.

Les résultats, présentés dans la figure , aux limites de certains seuils correspondent dans un cas (au dessous de 40%) à la détection des plus mauvais documents indexables et dans l'autre cas (au dessus de 70%) à la détection des meilleurs documents. Le seuil pourra être ajusté selon les compromis choisis entre coût et qualité d'indexation qui pourront être faits par un archiviste.

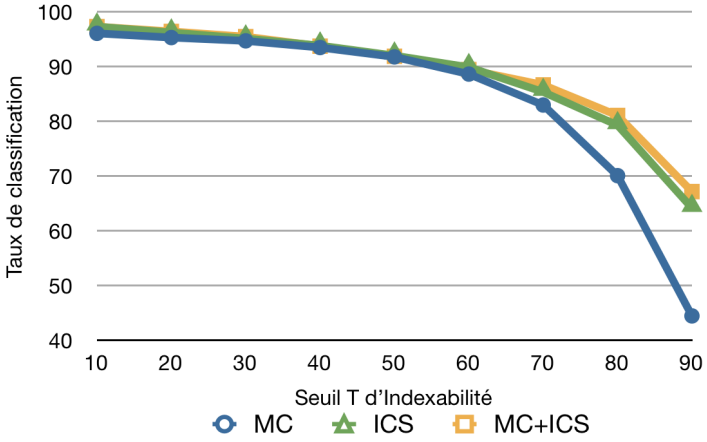

FIGURE 1 – Classification des documents en indexable ou non indexable selon un seuil de qualité qui varie de 10 à 90% d'indexabilité, en utilisant une prédiction de l'indexabilité des documents basée sur les mesures de confiance (MC), un index de compacité sémantique (ICS), une combinaison de MC et ICS ($MC + ICS$).

La mesure de confiance MC obtient de bonnes performances pour un seuil T d'indexabilité en dessous de 55. Effectivement, son taux de classification des documents nuisant à l'indexabilité est supérieur à 90%. Par contre, les performances chutent au-delà d'un seuil de 70 jusqu'à atteindre un taux de classification inférieur à 50%.

Par contre, la classification obtenue à l'aide d'une prédiction basée sur ICS est meilleure. Avec un seuil inférieur à 50, MC et ICS obtiennent sensiblement les mêmes taux de classification. Néanmoins, au delà du seuil de 70, la méthode basée sur la sémantique obtient de meilleurs résultats. Son taux de classification reste au dessus de 64% de classification (19,7% absolu)

La combinaison des deux méthodes améliore encore les résultats au dessus du seuil de 70 (gain moyen absolu supérieur à 2%). Ceci permet de valider que MC apporte une information non détenue par ICS.

Pour conclure, les résultats démontrent l'efficacité de la prédiction de l'indexabilité en utilisant plusieurs types de paramètres. Nos méthodes permettent de détecter avec fiabilité la qualité d'un segment de transcription en vue de son indexation. Cette prédiction permettra à un archiviste de valider les documents qui seront bien indexés et de détecter les documents qui seront mal indexés.

6 Conclusion et Perspectives

Dans cette étude, nous avons étudié l'intérêt de l'ajout d'information sémantique pour l'estimation de la qualité d'une transcription destinée à de la recherche d'information dans les documents audio. Nous avons introduit une méthode pour la prédiction de l'indexabilité qui combine une mesure de confiance issue du SRAP et un index de compacité sémantique. Les résultats montrent que l'information sémantique est un paramètre performant pour l'estimation des données destinées à la RIDA. Même si les résultats obtenus par la mesure de confiance issue du SRAP sont performants pour détecter les documents mal indexables, l'index de compacité sémantique permet d'obtenir une meilleure détection des documents correctement indexables, avec un gain moyen absolu supérieur à 10%. Ces résultats valident et améliorent les résultats obtenus dans l'étude précédente. Ceci s'expliquent principalement par un apprentissage des perceptrons sur un ensemble plus important de données. Nous envisageons maintenant d'étudier diverses stratégies de modélisation sémantique, en élargissant le contexte et le paradigme de modélisation (comme l'allocation latente de Dirichlet) qui pourrait améliorer l'extraction et la détection de concepts latents dans le flux de parole.

Références

Chelba, C., Hazen, T. et Saraclar, M. (2008). Retrieval and browsing of spoken content. *Signal Processing Magazine, IEEE*, 25(3):39–49.

Chelba, C., J. Silva et Acero, A. (2007). Soft indexing of speech content for search in spoken documents. *Computer Speech and Language*, 21:458–478.

Cox, S. et Dasmahapatra, S. (2002). High-level approaches to confidence estimation in speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 10(7):460–471.

Garofolo, J. S., Auzanne, C. G. P. et Voorhees, E. M. (2000). The TREC spoken document retrieval track : A success story. *In in TREC 8*, pages 16–19.

Hakkani-Tür, D., Tur, G., Ricardi, G. et Kim, H. K. (2005). Error prediction in spoken dialog : from signal-to-noise ratio to semantic confidence scores. volume I, pages 1041–1044.

Hansen, J., Huang, R., Zhou, B., Seadle, M., Deller, J., Gurijala, A., Kurimo, M. et Angkititrakul, P. (2005). Speechfind : Advances in spoken document retrieval for a national gallery of the spoken word. *Speech and Audio Processing, IEEE Transactions on*, 13(5):712–730.

Hatcher, E. et Gospodnetic, O. (2004). *Lucene in Action (In Action series)*. Manning Publications Co., Greenwich, CT, USA.

Kurimo, M. et Turunen, V. (2005). Retrieving speech correctly despite the recognition errors. *In 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*.

l. Chang, H., c. Pan, Y. et s. Lee, L. (2008). Latent semantic retrieval of spoken documents over position specific posterior lattices. *In SLT Workshop, 2008. SLT 2008. IEEE*, pages 285–288.

Moreno, P., Logan, B. et Raj, B. (2001). A boosting approach for confidence scoring. *In Interspeech, Aalborg, Denmark*, pages 2109–2112.

Oard, D. W., Soergel, D., Doermann, D., Huang, X., Murray, G. C., Wang, J., Ramabhadran, B., Franz, M., Gustman, S., Mayfield, J., Kharevych, L. et Strassel, S. (2004). Building an information retrieval test collection for spontaneous conversational speech. *In SIGIR '04*, pages 41–48, New York, USA. ACM.

ROSENBLATT, F. (1962). Principles of neurodynamics : Perceptrons and the theory of brain mechanisms. *In Spartan Books*.

SARACLAR, M. (2004). Lattice-based search for spoken utterance retrieval. *In In Proceedings of HLT-NAACL 2004*, pages 129–136.

SENAY, G., LINARÈS, G. et LECOUTEUX, B. (2011). A segment-level confidence measure for spoken document retrieval. *In ICASSP*, pages 5548–5551.

SIEGLER, M. A. (1999). *Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance*. Thèse de doctorat.

WHITTAKER, S., HIRSCHBERG, J., AMENTO, B., STARK, L., BACCHIANI, M., ISENHOUR, P. et GARY, S. (2002). Scanmail : a voicemail interface that makes speech browsable, readable and searchable. *In Proceedings of CHI2002*, pages 275–282. ACM Press.