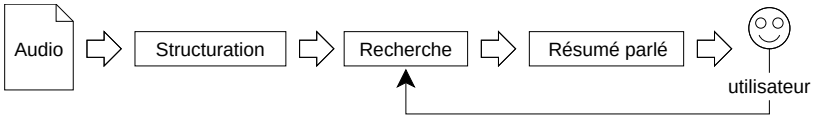
Accès aux connaissances orales par le résumé automatique

Benoît Favre [*,**] Jean-François Bonastre[**], Patrice Bellot[**], François Capman[*]

[*]Thales, Laboratoire MMP, 160 Bd de Valmy, 92700 Colombes,
francois.capman@fr.thalesgroup.com
[**]Université d'Avignon, LIA, 339 Ch des Meinajaries, 84000 Avignon,
benoit.favre@univ-avignon.fr
jean-francois.bonastre@univ-avignon.fr
patrice.bellot@univ-avignon.fr

Le temps nécessaire pour écouter un flux audio est un facteur réduisant l'accès efficace à de grandes archives de parole. Une première approche, la structuration automatique des données, permet d'utiliser un moteur de recherche pour cibler plus rapidement l'information. Les listes de résultats générées sont longues dans un souci d'exhaustivité. Alors que pour des documents textuels, un coup d'oeil discrimine un résultat interessant d'un résultat non pertinant, il faut écouter l'audio dans son intégralité pour en capturer le contenu. Nous proposons donc d'utiliser le résumé automatique afin de structurer les résultats des recherches et d'en réduire la redondance.



Les données radiophoniques exploitées pour cette approche sont issues de la campagne ESTER (Galliano et al., 2005), évaluatrice de la structuration automatique d'émissions et de bulletins à caractère informatif. Le processus de structuration de notre système est le suivant : segmentation en classes acoustiques (Fredouille et al., 2004), segmentation en locuteurs (Istrate et al., 2005), transcription de la parole (Nocera et al., 2004), segmentation thématique (Sitbon et Bellot, 2004), et reconnaissance d'entités nommées (Favre et al., 2005). Grâce à cette structuration, un moteur de recherche basé sur le modèle vectoriel permet de présenter à l'utilisateur la liste des segments correspondant à son besoin en information.

Fondé sur l'observation que 70% des phrases d'un résumé écrit manuellement proviennent des textes d'origines, le résumé par extraction est l'approche la plus utilisée actuellement en domaine ouvert pour le texte. En prenant pour hypothèse que cette observation est similaire pour la parole (les titres des journaux radiodiffusés), nous l'appliquons à la fois pour extraire des étiquettes thématiques structurant hiérarchiquement les résultats et pour extraire les segments les plus représentatifs du contenu des résultats.

L'algorithme *Maximal Marginal Relevance* (MMR), proposé par (Goldstein et al., 2000) pour sélectionner les segments maximisant la couverture en information tout en minimisant sa redondance, peut être appliqué pour sélectionner des mots-clés comme étiquettes thématiques dont on obtient une hiérarchie en faisant varier la granularité. Le critère de sélection par gain en

couverture de MMR est modifié en transposant le paradigme de représentation des documents par des vecteurs de mots, afin de représenter des mots par des vecteurs de documents.

$$\hat{t}_{i+1} = argmax_{t \notin sel} \lambda sim(\vec{t}, \vec{c_{res}}) - (1-\lambda) sim(\vec{t}, \vec{c_{sel}}) \quad (1)$$

Ici, $\vec{t}$ est le vecteur modélisant un mot-clé, $\vec{c_{res}}$ le vecteur centroïde des résultats, $\vec{c_{sel}}$ le vecteur centroïde de la sélection courante et $sim()$ la similarité mesurée par le cosinus de l'angle entre les vecteurs. Dans le domaine de l'information radiodiffusée, les mots-clés utilisés sont des entités nommées car les noms de lieux, de personnes et d'organisation permettent de caractériser des événements. Ces étiquettes thématiques sont proposées à l'utilisateur qui, en les sélectionnant, implique la restriction des résultats par conjonction avec les termes de la requête. Parallèlement, le résumé des segments audio est généré selon MMR classique pour permettre à l'utilisateur d'écouter l'équivalent d'un court bulletin d'informations.

Bien que le système permette une forte réduction du temps d'écoute, le résumé audio est soumis aux mêmes problèmes majeurs que le résumé textuel, à savoir les références non résolues et la réduction de redondance à l'interieur même des segments. S'ajoutent les erreurs de la structuration automatique et les désagréments liés à la parole comme les difficultés d'élocution ou les recouvrements de locuteurs dont l'impact est présent à l'écoute. Nous projetons pour la suite de ces travaux, d'adresser ces problèmatiques et d'évaluer le système d'accès aux flux de données parlées.

Références

Favre, B., F. Béchet, et P. Nocéra (2005). Robust named entity extraction from large spoken archives. In *HLT-EMNLP'05*.

Fredouille, C., D. Matrouf, G. Linares, et P. Nocera (2004). Segmentation en macro-classes acoustiques d'émissions radiophoniques dans le cadre d'ester. In *JEP'04*.

Galliano, S., E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, et G. Gravier (2005). The ESTER phase II evaluation campaign for the rich transcription of french broadcast news. In *Proc. Eurospeech'05*.

Goldstein, J., V. Mittal, J. Carbonell, et J. Callan (2000). Creating and evaluation multi-document sentence extract summaries. In *CIKM 2000 - ACM, McLean, VA USA*.

Istrate, D., N. Scheffer, C. Fredouille, et J.-F. Bonastre (2005). Broadcast news speaker tracking for ester 2005 campaign. In *Eurospeech'05*.

Nocera, P., C. Fredouille, G. Linares, D. Matrouf, S. Meignier, J.-F. Bonastre, D. Massonié, et F. Béchet (2004). The LIA's french broadcast news transcription system. In *SWIM*.

Sitbon, L. et P. Bellot (2004). Evaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au français. In *TALN 2004, Fès, Maroc*.

Summary

We propose to reduce listening time in spoken archives access interfaces : search engine results are structured according to automatically extracted concept hierachies and the rendondancy of results is removed using automatic summarization techniques.