UMR IRISA
INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTEMES ALEATOIRES
Campus universitaire de Beaulieu ● Avenue du Général Leclerc 35042 Rennes Cedex France
● www.irisa.fr

OFFRE de THÈSE

Sujet de thèse – PhD Subject Title:	**Résumés multidocuments multimédias**
Résumé – Abstract:	Une façon d'explorer une base de contenus multimédias (TV, vidéos, user generated contents (YouTube...), sites Web...) est de naviguer entre les documents disponibles en suivant, par exemple, un lien thématique. On peut ainsi prendre l'extrait d'un journal télévisé d'une chaîne de TV abordant un fait divers, obtenir de l'information sur le lieu où se déroule ce fait sur un site Web, retrouver dans des journaux d'autres chaînes TV la façon donc ce même fait divers a été présenté, voire explorer des réseaux sociaux à la recherche de commentaires émis sur le sujet. Si ce mode de navigation thématique est un moyen d'accès pratique pour un utilisateur, c'est actuellement à lui qu'incombe la tâche de synthétiser l'ensemble de l'information que sa navigation lui a permis de récolter, afin, malgré les redites nombreuses, de collecter l'ensemble des points-clés correspondant au sujet. Le sujet de cette thèse se situe dans ce cadre et a pour objectif d'aider l'utilisateur dans cette tâche en offrant des mécanismes de production de résumés multidocuments multimédias. Cet objectif final étant vaste, on se focalisera tout d'abord sur l'obtention d'un résumé à partir du média langage contenu dans les documents, que ce soit en exploitant des données textuelles écrites (sites Web, réseaux sociaux, documents pdf...) ou en accédant au texte prononcé dans des vidéos grâce à une transcription automatique de la parole. Pour ce faire, on fera dans un premier temps un état de l'art des techniques de résumés multidocuments issues du traitement automatique des langues. En effet, depuis quelques années, des travaux sont menés dans cette communauté sur des textes écrits bien formés. Ils permettent de produire des résumés multidocuments essentiellement par extraction des phrases saillantes en termes d'information contenue, et par fusion des phrases extraites en prenant soin de minimiser la redondance d'information inhérente à la méthodologie. La thèse visera entre autres l'étude de l'applicabilité de ces techniques à des mélanges de transcriptions de la parole (dépourvues de la notion de phrases et contenant des erreurs de transcription pouvant fausser l'extraction des phrases importantes) et de textes écrits. Dans un second temps, des indices issus des médias image, vidéo et son pourront être exploités, d'une part pour consolider l'émergence des parties saillantes à conserver pour le résumé, d'autre part pour étudier des possibilités de production de résumés sous une forme non uniquement langagière. Enfin, si des "benchmarks" d'évaluation de systèmes de résumés multidocuments textuels existent, la thèse devra également s'intéresser à la proposition d'une méthodologie d'évaluation des résumés produits dans ce cadre multimédia. Cette thèse s'inscrit pleinement dans les efforts de recherche actuels de l'équipe Texmex de l'IRISA sur l'utilisation de grandes collections multimédias et sur l'analyse de la composante orale de ces documents. Les travaux s'appuieront sur l'expertise développée dans Texmex autour de l'exploitation de la parole pour l'accès sémantique aux contenus multimédias.
Département scientifique – Scientific department:	D6 – Média et Interactions
Equipe projet -	TexMex ; http://www.irisa.fr/texmex/sujets

Research team:	
Directeur de thèse - PhD Director:	Pascale SÉBILLOT
Encadrant(s) – PhD supervisors:	Pascale SÉBILLOT Guillaume GRAVIER Patrick GROS
Contact(s) :	Pascale Sébillot pascale.sebillot@irisa.fr Guillaume Gravier guillaume.gravier@irisa.fr Patrick Gros patrick.gros@irisa.fr
Début des travaux - Work start date:	01/10/2013
Bibliographie - References:	R. Barzilay and K.R. McKeown (2005). Sentence fusion for multidocument news summarization. In Computational Linguistics 31(3): 297-328. K. Filippova (2010). Multi-sentence compression: Finding shortest paths in word graphs. In Proc. 23rd International Conference on Computational Linguistics (COLING 2010), pp. 322-330. J. Goldstein, V. Mittal, J. Carbonell and M. Kantrowitz (2000). Multi-document summarization by sentence extraction. In Proc. ANLP/NAACL Workshop on Automatic Summarization, pp. 40-48. J. Steinberger and K. Jezek (2009). Text summarization: An old challenge and new approaches, In Foundations of Computational Intelligence 6(SCI 206):127-149.