



MINISTÈRE DE L'ÉDUCATION
NATIONALE, DE L'ENSEIGNEMENT
SUPÉRIEUR ET DE LA RECHERCHE

>>>

AVIS DE SOUTENANCE DE THESE DE DOCTORAT

Monsieur Florian BOUDIN soutiendra une thèse
le 5 décembre 2008 à 14h

L.I.A.

Amphi B. Pascal

SPÉCIALITÉ : INFORMATIQUE ED 166

Titre de la thèse : exploration d'approches statistiques pour le résumé automatique de texte.

Membres du jury :

TORRES-MORENO Juan-Manuel, MCF- HDR Informatique, Université d'Avignon et des Pays de
Vaucluse

EL-BÈZE Marc, PR Informatique, Université d'Avignon et des Pays de Vaucluse

LAPALME Guy, PR Informatique, Université de Montréal, Canada,

SAGGION Horacio, Research Fellow, Informatique Université de Sheffield, Royaume Uni

GALLINARI Patrick, PR Informatique, Université de Paris 6,

POIBEAU Thierry, MCF Informatique, Université de Paris 1

Résumé de la thèse :

Un résumé est un texte reformulé dans un espace plus réduit.

Il doit exprimer avec un minimum de mots le contenu essentiel d'un document.

Son but est d'aider le lecteur à repérer les informations qui peuvent l'intéresser sans pour autant
devoir lire le document en entier. Mais pourquoi avons-nous tant besoin de résumés ? Simplement
parce que nous ne disposons pas d'assez de temps et d'énergie pour tout lire. La masse
d'information textuelle sous forme électronique ne cesse d'augmenter, que ce soit sur Internet ou
dans les réseaux des entreprises.

Ce volume croissant de textes disponibles rend difficile l'accès à l'information désirée sans l'aide
d'outils spécifiques. Produire un résumé est une tâche très complexe car elle nécessite des
connaissances linguistiques ainsi que des connaissances du monde qui restent très difficiles à
incorporer dans un système automatique.

Dans cette thèse de doctorat, nous explorons la problématique du résumé automatique par le biais
de trois méthodes statistiques permettant chacune la production de résumés répondant à une tâche
différente.

Nous proposons une première approche pour la production de résumé dans le domaine spécialisé de
la Chimie Organique. Un prototype nommé YACHS a été développé pour démontrer la viabilité de
notre approche. Ce système est composé de deux modules, le premier applique un pré-traitement
linguistique particulier afin de tenir compte de la spécificité des documents de Chimie Organique
tandis que le second sélectionne et assemble les phrases à partir de critères statistiques dont certains
sont spécifiques au domaine.

Nous proposons ensuite une approche répondant à la problématique du résumé automatique multi-
documents orienté par une thématique.

Nous détaillons les adaptations apportées au système de résumé générique Cortex ainsi que les
résultats observés sur les données des campagnes d'évaluation DUC.

Les résultats obtenus par la soumission du LIA lors des participations aux campagnes d'évaluations
DUC 2006 et DUC 2007 sont discutés.

Nous proposons finalement deux méthodes pour la génération de résumés "mis-à-jour".

La première approche dite de maximisation-minimisation a été évaluée par une participation à la
tâche pilote de DUC 2007.

La seconde méthode est inspirée de Maximal Marginal Relevance (MMR), elle a été évaluée par
plusieurs soumissions lors de la campagne TAC 2008.