THESE / UNIVERSITE DE BRETAGNE-SUD
sous le sceau de l'Université Bretagne Loire

pour obtenir le titre de
DOCTEUR DE L'UNIVERSITE DE BRETAGNE-SUD

Mention : Informatique
Ecole doctorale SICMA

Présentée par
VU Hai Hieu

Préparée dans l'équipe EXPRESSION
Laboratoire IRISA

Indexation aléatoire et similarité inter-phrases appliquées au résumé automatique

Thèse soutenue le 29 janvier 2016

devant le jury composé de :

Pierre-François MARTEAU
Professeur, université de Bretagne Sud / directeur de thèse

Jeanne VILLANEAU
MCF, université de Bretagne Sud / co-directrice de thèse

Farida SAÏD
MCF, université de Bretagne Sud / co-directrice de thèse

Sophie ROSSET
Chercheuse, LIMSI – CNRS / rapporteuse

Emmanuel MORIN
Professeur, université de Nantes / rapporteur

Gwénolé LECORVÉ
MCF, université de Rennes 1 / examinateur

UNIVERSITE DE BRETANGE-SUD

Résumé

IRISA
EXPRESSION

Docteur en informatique

Indexation aléatoire et similarité inter-phrases appliquées au résumé automatique

par VU Hai Hieu

Face à la masse grandissante des données textuelles présentes sur le Web, le résumé automatique d'une collection de documents traitant d'un sujet particulier est devenu un champ de recherche important du Traitement Automatique des Langues. Les expérimentations décrites dans cette thèse s'inscrivent dans cette perspective. L'évaluation de la similarité sémantique entre phrases est l'élément central des travaux réalisés. Notre approche repose sur la similarité distributionnelle et une vectorisation des termes qui utilise l'encyclopédie Wikipédia comme corpus de référence. Sur la base de cette représentation, nous avons proposé, évalué et comparé plusieurs mesures de similarité textuelle ; les données de tests utilisées sont celles du défi SemEval 2014 pour la langue anglaise et des ressources que nous avons construites pour la langue française. Les bonnes performances des mesures proposées nous ont amenés à les utiliser dans une tâche de résumé multi-documents, qui met en oeuvre un algorithme de type PageRank. Le système a été évalué sur les données de DUC 2007 pour l'anglais et le corpus RPM2 pour le français. Les résultats obtenus par cette approche simple, robuste et basée sur une ressource aisément disponible dans de nombreuses langues, se sont avérés très encourageants.

Remerciements

Je tiens à remercier, en tout premier lieu, mon directeur et mes co-directeurs de thèse, Monsieur le Professeur Pierre-François MARTEAU, Mesdames Jeanne VILLANEAU et Farida SAÏD pour m'avoir accueilli, guidé et mis dans les meilleures conditions pour préparer ma thèse au sein de l'équipe EXPRESSION du Laboratoire IRISA, l'Université de Bretagne-Sud. Je tiens à leur exprimer ma gratitude pour leurs qualités pédagogiques et scientifiques, leur franchise, leur sympathie, leur confiance. J'ai appris beaucoup auprès d'eux. Je leur suis également reconnaissant pour leur écoute, leur partage et leur soutien dans les moments difficiles. J'ai pris un grand plaisir à travailler sous leur direction.

Je voudrais aussi remercier les rapporteurs de cette thèse : Madame Sophie ROSSET, Directrice de Recherche du Laboratoire LIMSI, CNRS et Monsieur le Professeur Emmanuel MORIN au Laboratoire d'Informatique de Nantes-Atlantique, l'Université de Nantes pour l'intérêt qu'ils ont porté à mon travail.

Mes remerciements s'adressent également à Monsieur Gwénolé LECORVÉ de l'Université de Rennes 1 pour avoir accepté d'examiner mon travail et de participer au jury.

Je souhaite remercier tous les membres du laboratoire IRISA, Lab-STICC, ENSIBS : les enseignants, techniciens, administratifs et doctorants qui m'ont aidé et accompagné dans mon travail durant ces quatre années en France.

Je n'oublie pas non plus tous les amis de France qui nous ont aidés, ma famille et moi : Brigitte ENQUEHARD, Evelyne BOUDOU, Alain BOUDOU, Lucien MOREL, Gildas TRÉGUIER, Sylvain CAILLIBOT..., les étudiants vietnamiens et les familles vietnamiennes de Lorient.

Pour terminer, je remercie du fond du cœur mes beaux-parents NONG Quoc Chinh - TRAN Thi Doan, mes parents VU The Huan - LE Thi Nhi et tous les membres de ma famille qui m'ont toujours soutenu, tout au long de ma vie, de mes études, sans lesquels je n'en serais pas là aujourd'hui. Ma reconnaissance va surtout à mon épouse NONG Thi Quynh Tram et à nos deux enfants VU Quynh Maï et VU Haï Minh qui sont toujours à mes côtés et me donnent la force de relever les défis.

Table des matières

Table des figures