COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Maël HOUBRE

Génération de mots-clés absents pour l'indexation d'articles scientifiques

Thèse présentée et soutenue à Nantes, le 10 Juillet 2025
Unité de recherche : UMR6004 – Laboratoire des Sciences et du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Mathieu CONSTANT	Professeur, Université de Lorraine (ATILF)
Davide BUSCALDI	Maître de conférences, Université Sorbonne Paris Nord (LIPN)

Composition du Jury :

Président :		
Examinateurs :	Lynda TAMINE-LECHANI	Professeure, Université Paul Sabatier (IRIT)
	Liana ERMAKOVA	Maître de conférences, Université de Bretagne Occidentale (HCTI Brest)
	Mathieu CONSTANT	Professeur, Université de Lorraine (ATILF)
	Davide BUSCALDI	Maître de conférences, Université Sorbonne Paris Nord (LIPN)
Dir. de thèse :	Béatrice DAILLE	Professeure, Université de Nantes (LS2N)
Co-encadrant :	Florian BOUDIN	Maitre de conférence, Université de Nantes (LS2N)

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

Titre : Génération de Mots-Clés Absents pour l'Indexation d'Articles Scientifiques

Mot clés : génération de mots-clés absents, indexation d'articles scientifiques, augmentation de données, homogénéité de l'indexation

Résumé : Les méthodes récentes de prédiction de mots-clés peuvent générer des mots-clés qui n'apparaissent pas dans le texte du document. Ces mots-clés "absents" apportent de nouvelles informations sur les documents et améliorent ainsi l'indexation. Néanmoins, les performances des modèles sur la génération de mots-clés absents sont faibles. Dans cette thèse, nous nous intéressons à l'amélioration de ces performances. Dans un premier temps, nous introduisons KPbiomed, un grand jeu de données dans le domaine biomédical. Il nous permet d'étudier l'influence de la quantité de données d'entraînement sur les performances. Nos résultats montrent qu'utiliser plus de données d'entraînement améliore les performances. L'obtention d'une telle quantité de données peut être difficile. Nous présentons dans un second temps une méthode d'augmentation de données basée sur la recombinaison de documents partageant des mots-clés. Nos résultats illustrent l'amélioration des performances de génération grâce à nos exemples synthétiques. Augmenter les performances sur les mots-clés absents suggère une meilleure capacité à relier un document à un concept abstrait. Nous présentons dans un troisième temps une méthode d'évaluation de l'homogénéité de l'indexation i.e. la faculté de prédire le même mot-clé pour différents documents traitant d'un même concept. Nos évaluations soulignent l'impact négatif des performances génératives sur l'homogénéité des modèles.

Title: Absent Keyphrase Generation for Scientific Indexing

Keywords: absent keyphrases generation, scientific indexing, data augmentation, indexing homogeneity

Abstract: Recent keyphrase prediction methods can predict keyphrases that do not appear in the source text. Those "absent" keyphrases improve the indexation by bringing new information on documents. However, performances on thoses absent keyphrases are low. This thesis focuses on improving those performances. First, we introduce KPBiomed, a large scale dataset in the biomedical domain. It allows us to measure the impact of training data on performances. Results show that using more training data improves performances. Obtaining such quantity of data can be difficult. We then present a data augmentation method. Experiments show improvements in performances with our artificial samples based on documents sharing keyphrases. Better absent keyphrase generation performances suggest a better ability to link a document to an abstract concept. We then present a method to evaluate homogeneity i.e. the capacity to predict the same keyphrase for different documents on the same concept. Results show to our surprise that absent keyphrase generation performances have a negative impact on homogeneity.