REDIRE : Réduction Extrême de DImension pour le Résumé Extractif

Marius Ortega*, Aurélien Bossard,** Nédra Mellouli,*,** Christophe Rodrigues*

*Léonard De Vinci Pôle Universitaire, Research Center, 92916 Paris La Défense, France
**Laboratoire d'Intelligence Artificielle et Sémantique des Données,
Université Paris 8 (EA4383), 93200 Saint-Denis, France
marius.ortega@edu.devinci.fr; nedra.mellouli@devinci.fr;
aurelien.bossard@iut.univ-paris8.fr; christophe.rodrigues@devinci.fr

Résumé. Nous présentons un modèle de résumé automatique non supervisé capable d'extraire les phrases les plus importantes d'un ensemble de textes. Pour extraire les phrases dans un résumé, nous utilisons des plongements de mots pré-entraînés afin de représenter les documents. A partir de cet épais nuage de vecteurs de mots, nous appliquons une réduction extrême de dimension permettant d'identifier des mots importants que nous regroupons par proximité. Les phrases sont extraites grâce à l'optimisation linéaire pour maximiser l'information présente dans le résumé. Nous évaluons l'approche sur des documents de grande taille et présentons des premiers résultats très encourageants.

1 Introduction

Dans cet article, nous présentons une méthode de Réduction Extrême de DImension pour le Résumé Extractif (REDIRE). Cette méthode est entièrement non supervisée. Si les méthodes génératives supervisées ont vu le plus de progrès notables depuis *Pointer-Generator* See et al. (2017), le premier système de résumé génératif par apprentissage profond, les méthodes extractives restent nécessaires pour s'affranchir des corpus d'apprentissage imposants et coûteux, des limites de domaine et des limites liées à la taille des documents en entrée.

L'idée de cette méthode nous est venue d'une étude sur les plongements de mots, que nous présentons en Section 3. Nous décrivons la méthode en Section 4 avant de montrer en Section 5 qu'elle s'applique avec de bons résultats à des documents longs.

2 Travaux connexes

TextRank Mihalcea et Tarau (2004) est sûrement la *baseline* la plus utilisée en résumé automatique non supervisé. Elle sélectionne les phrases des documents source d'après leur centralité dans une représentation graphique du document fondée sur les similarités entre phrases, grâce à un algorithme de marche aléatoire dans le graphe. Gillick et Favre (2009) considèrent le résumé comme un problème d'optimisation de sac à dos visant à maximiser la somme des