Extension des mesures textuelles d'informativité à l'évaluation de l'intérêt potentiel d'un passage

Carlos E. González-Gallardo*, Éric SanJuan-Ibekwe*, Juan Manuel Torres Moreno*

*Laboratoire d'Informatique d'Avignon
339 chemin des Meinajaries, 84911 Avignon cedex 9, FRANCE
eric.sanjuan@univ-avignon.fr
http://lia.univ-avignon.fr/

1 Introduction

Les mesures d'informativités utilisées pour évaluer le résumé automatique de textes s'appuient pour la plupart sur des mesures de recouvrement entre les n-grammes présents dans le résumé produit automatiquement et ceux apparaissant dans un résumé de référence (Torres-Moreno, 2014) généralement produit par un expert. Ces mesures diffèrent selon :

— la métrique utilisée (cosinus, Dice, Rouge, Kullback-Leibler, Similarité Logarithmique)
— le sac des termes considéré (mots simples, mots n -grammes, entités, pépites, etc.).

Récemment, les approches par plongement lexical de mots offrent une alternative numérique à ces approches discrètes basées sur la présence / absence d'une unité de texte (Ng et Abrecht, 2015).

Ces mesures ont été ensuite étendues à l'évaluation de la recherche ciblée d'information par des requêtes complexes (Bellot et al., 2016). En particulier, dans la tâche INEX de contextualisation de tweets, ce sont des contenus entiers de mictroblogs qui ont été considérés comme des requêtes.

2 Proposition

Nous définissons formellement l'évaluation de l'informativité sur de courts passages comme une relation ternaire entre un ensemble de sujets T, un sous-ensemble d'extraits courts P provenant d'une grande ressource documentaire et un ensemble S de notes normalisées tel que les passages classés en tête contiennent le plus d'informations les plus pertinentes vis à vis des thèmes explicitement mentionnées dans le passage ou sur des sujets connexes liés de manière implicite.

Nous définissons alors l'extension du concept d'intérêt (Koh et al., 2008) appliqué au texte comme une généralisation de la notion d'informativité où la requête traduisant un besoin d'information de l'utilisateur serait inconnue a priori mais serait susceptible d'être explicité à posteriori.