Représentations lexicales pour la détection non supervisée d'événements dans un flux de tweets : étude sur des corpus français et anglais

Béatrice Mazoyer*,**, Nicolas Hervé**,
Céline Hudelot*, Julia Cagé***

*CentraleSupélec (Université Paris-Saclay), MICS, Gif-sur-Yvette, France
beatrice.mazoyer, celine.hudelot@centralesupelec.fr,
**Institut National de l'Audiovisuel, Bry-sur-Marne, France
nherve@ina.fr
***SciencesPo Paris, Département d'économie, Paris, France
julia.cage@sciencespo.fr

Résumé. Dans cet article, nous nous intéressons aux approches récentes de plongements lexicaux en vue de les appliquer à la détection automatique d'événements dans un flux de tweets. Nous modélisons cette tâche comme un problème de clustering dynamique. Nos expériences sont menées sur un corpus de tweets en anglais accessible publiquement ainsi que sur un jeu de données similaire en français annoté par notre équipe. Nous montrons que les techniques récentes fondées sur des réseaux de neurones profonds (ELMo, Universal Sentence Encoder, BERT, SBERT), bien que prometteuses sur de nombreuses applications, sont peu adaptées pour cette tâche, même sur le corpus en anglais. Nous expérimentons également différents types de fine-tuning afin d'améliorer les résultats de ces modèles sur les données en français. Nous proposons enfin une analyse fine des résultats obtenus montrant la supériorité des approches traditionnelles de type tf-idf pour ce type de tâche et de corpus.

1 Introduction

Les recherches récentes en traitement automatique du langage ont permis d'atteindre des performances proches des capacités humaines, notamment en ce qui concerne la détection de paraphrase ou l'évaluation de la similarité sémantique entre deux phrases [1]. Cependant, ces avancées, fondées sur l'entraînement de réseaux de neurones sur de très vastes corpus de textes, sont à nuancer.

En effet, malgré des progrès rapides ces dernières années dans l'adaptabilité des modèles de traitement du langage (GLUE, le benchmark de référence (Wang et al., 2018), est constitué de 9 tâches différentes, et les modèles sont évalués en fonction de leur performance moyenne sur toutes ces tâches), il reste difficile d'adapter ces modèles à de nouvelles tâches. Dans cet article,

1. Voir les résultats obtenus sur le benchmark GLUE : `gluebenchmark.com`