Construction automatique de lexiques thématiques

Suzanne Mpouli*, Christine Largeron*, Michel Beigbeder**

*UJM-Saint-Etienne, Laboratoire Hubert Curien UMR 5516, F-42023, France
Suzanne.Mpouli@univ-st-etienne.fr, Christine.Largeron@univ-st-etienne.fr

**Univ Lyon, IMT Mines Saint-Etienne, Institut Henri Fayol,
Univ Jean Monnet, IOGS, CNRS, LHC, F - 42023 Saint-Etienne FRANCE
michel.beigbeder@emse.fr

Résumé. Nous présentons *Lexifield*, un système entièrement automatique et indépendant de la langue pour la construction de lexiques spécifiques à un domaine à partir de courtes listes de termes¹. *Lexifield* s’appuie sur un modèle de plongement lexical (*word embedding*), un dictionnaire de définitions et un dictionnaire de synonymes. Pour évaluer *Lexifield*, quatre lexiques ont été générés : un lexique en français pour le sujet “son” et trois lexiques en anglais pour les sujets “sound”, “taste” et “odour”. Les résultats ont confirmé ses bonnes performances par rapport à d’autres systèmes de l’état de l’art.

1 Introduction

La création manuelle de lexiques est non seulement longue et coûteuse mais peut également aboutir à un contenu qui n’est pas à jour. Notre objectif est donc de construire automatiquement des lexiques, au sens original du terme, puisque nous voulons générer des listes de mots qui couvrent un sujet spécifique : trois des cinq sens humains (odorat, ouïe et goût) dans notre étude de cas, bien que notre méthode puisse également être appliquée à d’autres sujets.

Le même problème a été abordé par Tekiroglu et al. (2014) qui génère Sensicon, un lexique se rapportant aux cinq sens à partir d’une liste de 277 lexèmes sélectionnés manuellement dans FrameNet (Baker et al. (1998)) et étendue automatiquement avec WordNet (Fellbaum (1998)). Enfin, les voisins des mots appartenant à la liste filtrée sont récupérés dans un grand corpus. Ces mots candidats sont classés à l’aide de la mesure NPMI (*Normalized Pointwise Mutual Information*) pour obtenir une liste finale triée.

Une approche similaire est décrite dans Riloff et Shepherd (1997) qui utilise uniquement une courte liste de mots relatifs au sujet sélectionné et un corpus lié au domaine pour étendre celle-ci. Les mots qui co-apparaissent dans une phrase avec les mots initiaux sont ordonnés en fonction de leur fréquence d’occurrence avec les mots d’origine et de leur fréquence d’apparition dans l’ensemble du corpus. Ce système a été amélioré par Roark et Charniak (1998) où la connaissance syntaxique est utilisée pour éliminer du bruit dans la liste des résultats. Le système Empath (Fast et al. (2016)) est similaire, mais utilise des technologies plus récentes telles qu’un modèle de plongement de mots entraîné sur le corpus Wattpad² et des outils de

1. Ce travail a été réalisé dans le cadre du projet SoundCITYve soutenu par le Labex IMU

2. <http://wattpad.com>