

Expression des connaissances en langage naturel : singularité et normalité d'une sélection

Jérémy Vizzini*, Cyril Labbé*, François Portet*

* Univ. Grenoble Alpes, CNRS, LIG, F-38000 Grenoble France
nom.prenom@imag.fr,

1 Expression des connaissances en langage naturel

La richesse du langage naturel permet de résumer des informations complexes et nombreuses en les rendant accessibles à tous. Un texte peut être écouté par des personnes malvoyantes et adapté à l'expertise du destinataire. Le domaine de la génération automatique de textes (GAT) offre donc des perspectives pertinentes et intéressantes pour transmettre des connaissances riches, complexes et personnalisées.

Un exemple d'application de la GAT pour la transmission d'information à grande échelle est le journalisme automatisé. Ce domaine d'application a connu un fort engouement ces dernières années. Par exemple, on peut citer Syllabs (2016) dont la solution `Data2Content` a été utilisée pour publier des billets de résultats d'élections sur le site `lemonde.fr`. Cependant, on peut constater que la plupart des textes générés restent très descriptifs des données en entrée se limitant au cercle restreint de l'entité à décrire. Cependant, pour transmettre de l'information, il est pertinent de la mettre en perspective avec d'autres informations par exemple en signalant des rapports avec des informations semblables ou en exprimant des similarités et différences notables. On peut ainsi mentionner des évolutions ou des corrélations en rapport avec les données de l'entité décrite mais n'en faisant pas partie explicitement. Par exemple les prévisions météo ou les résultats d'élections concernant une localité peuvent être comparés aux données concernant des localités ayant des propriétés identiques, p.ex., de la même région etc. Informations qui peuvent être ensuite insérées dans le texte.

Nous présentons le prototype `Summy` qui est un outil permettant de construire un générateur de textes et offrant la possibilité de transcrire en langage naturel les singularités et/ou la normalité d'un ensemble de données. La démarche consiste à identifier et expliciter les ressources et connaissances (modèles, ressources langagières etc.) nécessaires à la production du générateur de textes. L'objectif est de rendre l'approche générique et applicable à moindre coût dans différents domaines d'application (météo, élections, sports...). Le prototype a été testé avec des données d'élections régionales.

2 Enrichir les textes décrivant des résultats d'élections

Pour les sites d'information, le défi réside en la génération en un temps minimal d'un grand nombre de textes présentant les résultats individualisés de chaque zone de vote. Pour des