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

Jérémy Vizzini, Cyril Labbé, François Portet

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 language 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

élections de grande ampleur il est bien trop coûteux de faire réaliser le travail de rédaction par l'homme. Des sites d'information ont mis en place des systèmes de génération automatique de textes dont l'exemple ci-dessous est extrait.

Exemple (texte des résultats des élections avec enrichissements)**.** *Au second tour des élections régionales 2015, la région Auvergne-Rhône-Alpes voyait trois candidats s'affronter : M. Jean-Jack Queyranne (liste Union de la Gauche), M. Laurent Wauquiez (liste Union de la Droite) et M. Christophe Boudot (liste Front National).*
La liste Union de la Gauche est arrivée en première position avec 57,19 % des voix. Elle a devancé la liste Union de la Droite qui a obtenu 29,78 % et la liste Front National, qui a recueilli 13,03 % des voix.
...

Cependant, la base de données [1] contient les connaissances nécessaires à la génération d'un texte contenant d'autres informations soulignant le caractère *normal* ou *exceptionnel* des données choisies tels que : *Les résultats de la ville de Grenoble sont* très similaires *à ceux des autres communes de la région.* ou *Cette tendance à voter à Gauche est* habituelle *pour une ville ayant une population jeune aux revenus moyens*. Le prototype `Summy` permet de construire des générateurs de textes qui incluent des enrichissements de ce type.

3 Conclusion

Le prototype `Summy` est un outil pour construire des générateurs de textes permettant d'inclure les enrichissements présentés ci-dessus. L'approche Labbé et al. (2015); Labbé et Portet (2012) permet de résumer et de contextualiser en langage naturel les données résultats d'une requête utilisateur. Cette approche a été testée sur des données de résultats d'élection. La prochaine étape consiste à mettre en place une évaluation sur un ensemble plus diversifié de données (sport, bourse, etc.) pour mesurer les performances techniques (temps d'exécution, qualité linguistique et cohérence du texte généré) et évaluer la préférence des lecteurs lors d'une étude comparative. Il conviendra aussi de mesurer le temps de développement nécessaire aux utilisateurs-développeurs pour créer de nouveaux générateurs.

Références

Labbé, C. et F. Portet (2012). Towards an abstractive opinion summarisation of multiple reviews in the tourism domain. In *The First International Workshop on Sentiment Discovery from Affective Data (SDAD 2012)*, pp. 87–94.

Labbé, C., C. Roncancio, et D. Bras (2015). A Personal Storytelling about Your Favorite Data. In *15th European Workshop on Natural Language Generation (ENLG 2015)*, Brighton, United Kingdom.

Syllabs (2016). Nos robots rédacteurs collaborent avec le monde. http ://blog.syllabs.com/le-monde-elections-departementales-syllabs-robotjournalisme/. Accessed : 2016-10-14.

1. Pour les régions, les départements et communes résultats disponibles : `https://www.data.gouv.fr/fr/datasets/selection-thematique-elections-regionales-2015/`