Un outil modulaire libre pour le résumé automatique

Valentin Nyzam Aurélien Bossard
LIASD, Université Paris 8 - IUT de Montreuil, 140 rue de la Nouvelle France,
93100 Montreuil, France
valentin.nyzam@iut.univ-paris8.fr,
aurelien.bossard@iut.univ-paris8.fr

RÉSUMÉ

Nous proposons une démonstration d'un outil modulaire et évolutif de résumé automatique qui implémente trois méthodes d'extraction de phrases de l'état de l'art ainsi que sept méthodes d'évaluation des phrases. L'outil est développé en Java et est d'ores-et-déjà disponible sur la plateforme Github.

ABSTRACT

A Modular Open Source Tool for Automatic Summarization

We propose a demonstration of an evolutive and modular open source tool for automatic summarization. The tool is developed in Java, implements three sentence extraction methods and seven sentence scoring methods, and is available on Github platform.

MOTS-CLÉS : résumé automatique, open source.

KEYWORDS: automatic summarization, open source.

Le résumé automatique est une tâche explorée depuis les années 1950 (Luhn, 1958). Depuis, de nombreuses méthodes ont vu le jour, qui s'appuient sur des domaines aussi variés que l'analyse de graphes (Erkan & Radev, 2004; Mihalcea & Tarau, 2004), les modèles statistiques génératifs (Blei *et al.*, 2003), la programmation linéaire en nombres entiers (Gillick & Favre, 2009b) ou encore les algorithmes évolutionnaires (Bossard & Rodrigues, 2015). On peut être amené, pour des besoins d'évaluation ou d'expérimentation sur de nouveaux corpus, à tester l'efficacité de différentes méthodes reconnues. Cependant, l'implémentation de ces méthodes n'est pas toujours disponible. De plus, quand elle l'est, il est souvent compliqué de la comparer à d'autres, car elles mettent en œuvre des pré et post-traitements différents qui influent sur la qualité des résumés produits.

Certains systèmes existent déjà, par exemple MEAD [1], News In Essence ou encore ICSISUMM [2], mais ceux-ci implémentent uniquement la ou les méthodes de leurs auteurs. Sumy [3], un système de résumé automatique multidocument, implémente plusieurs méthodes de résumé, mais ne propose pas une approche modulaire qui permet de modifier chaque composant de la chaîne de traitement du résumé automatique. Au contraire, l'outil que nous proposons est le plus modulaire possible afin de permettre par exemple de tester l'apport de nouvelles approches du TAL en changeant la représentation des mots ou des phrases dans des méthodes de résumé déjà existantes.

L'objectif de la démonstration est de montrer l'utilisation et la modularité de ce nouvel outil open

1. http://www.summarization.com/mead/
2. https://code.google.com/archive/p/icsisumm/
3. https ://github.com/miso-belica/sumy

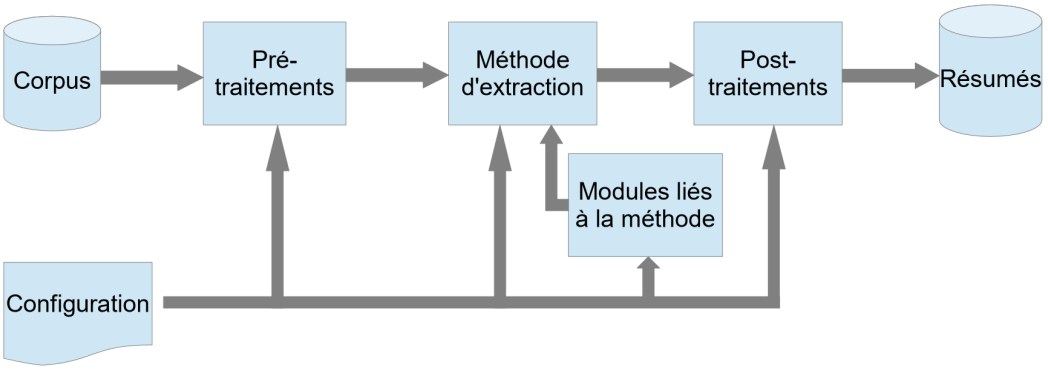

FIGURE 1 – Architecture globale de l'outil de résumé automatique

source évolutif pour le résumé automatique. Développé en Java, nous y avons implémenté des méthodes bien connues d'extraction de phrases pour le résumé automatique :

— MMR (Carbonell & Goldstein, 1998) ;
— ILP (Gillick & Favre, 2009b) ;
— Algorithme évolutionnaire (Bossard & Rodrigues, 2015).

Toutes ces méthodes d'extraction comportent des composants modulables. Par exemple, la méthode MMR nécessite des méthodes d'évaluation de la pertinence des phrases. Ainsi, nous avons implémenté les méthodes d'évaluation suivantes :

— Variantes fondées sur le tfidf ;
— Centroid (Radev *et al.*, 2004) ;
— LexRank (Erkan & Radev, 2004) ;
— LDA (Blei *et al.*, 2003) ;
— *Word embeddings* (Zhang *et al.*, 2015)

L'outil permet de choisir des pré-traitements, une méthode d'indexation, une méthode d'extraction ainsi que différentes options propres à cette méthode, et des post-traitements. Les méthodes communiquent entre elles leurs résultats à l'aide d'interfaces génériques. Cela permet ainsi d'ajouter une nouvelle méthode sans toucher au cœur de l'outil mais simplement en utilisant les interfaces proposées. L'outil dispose également d'un algorithme génétique (à bien différencier de l'algorithme évolutionnaire dédié au résumé) qui permet d'optimiser les paramètres de l'outil d'une méthode donnée sur un corpus pour lequel on dispose de résumés de référence.

Sans rentrer ici dans les détails de l'architecture de chacune de ces méthodes, l'architecture globale de l'outil est présentée en Figure 1.

Cet outil permet d'évaluer différentes méthodes dans un cadre unifié (pré et post traitements identiques) et nous espérons qu'il sera rapidement adopté par la communauté du résumé automatique afin qu'y soient implémentées de nouvelles méthodes de résumé automatique.

Un *web service* qui utilise l'outil de résumé dont nous ferons la démonstration est en cours de développement pour permettre d'obtenir rapidement, et sans avoir à l'installer, des résumés multidocuments.

1 Remerciements

Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet Jeunes Chercheurs/Jeunes Chercheuses ASADERA - convention ANR-16-CE38-0008)

Références

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

BOSSARD A. & RODRIGUES C. (2015). Une approche évolutionnaire pour le résumé automatique. In *TALN 2015 - 22ème Conférence sur le Traitement Automatique des Langues Naturelles*.

CARBONELL J. & GOLDSTEIN J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR'98 : Proceedings of the 21st ACM SIGIR Conference*, p. 335–336.

ERKAN G. & RADEV D. R. (2004). Lexrank : Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.

GILLICK D. & FAVRE B. (2009b). A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, p. 10–18 : Association for Computational Linguistics.

LUHN H. (1958). The automatic creation of literature abstracts. *IBM Journal*, **2**(2), 159–165.

MIHALCEA R. & TARAU P. (2004). TextRank : Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*.

RADEV D. R., JING H., STY M. & TAM D. (2004). Centroid-based summarization of multiple documents. *Information Processing Management*, **40**, 919–938.

ZHANG Y., ER M. J. & ZHAO R. (2015). Multi-document extractive summarization using window-based sentence representation. In *2015 IEEE Symposium Series on Computational Intelligence*, p. 404–410.