Juan-Manuel Torres-Moreno
INFORMATICIEN,
UNIVERSITÉ D'AVIGNON
Maître de conférences, spécialiste du traitement automatique des langues.

Sabine Louët
JOURNALISTE ET
ENTREPRENEUSE
Fondatrice de la société d'édition SciencePOD, basée à Dublin (Irlande).

L'intelligence artificielle a lu pour vous

Résumer automatiquement des publications scientifiques afin de permettre au plus grand nombre d'y avoir accès. Tel est l'objectif ambitieux d'une collaboration associant les connaissances en matière de résumé automatique de Juan-Manuel Torres-Moreno, du Laboratoire informatique d'Avignon (LIA), et le savoir-faire en communication scientifique de Sabine Louët, fondatrice de la société d'édition numérique SciencePOD. Une manière de diffuser les progrès de la recherche à grande échelle, et aussi de lutter contre la désinformation.

Comment se tenir rapidement informé des connaissances scientifiques dans un monde où prolifèrent les informations liées à la recherche ? Rappelons que le nombre d'études scientifiques publiées annuellement dans le monde est passé de 972 000 en 1996 à 2,5 millions en 2018, selon des chiffres de la National Science Foundation, aux États-Unis. Or tous ces travaux ne proposent pas forcément de résumé (ou abstract). Et, s'il existe, celui-ci contient souvent trop de termes inaccessibles à un public non averti. Dans ce contexte, la possibilité de recourir à l'intelligence artificielle (IA) pour générer des résumés automatiques montre toute sa pertinence.

Cette ambition de résumer des textes ne date pas de l'édition numérique. L'essayiste français Joseph Joubert (1754-1824) exprimait déjà cette volonté dans ses *Pensées* : *« S'il est un homme tourmenté par la maudite ambition de mettre tout*

JEAN-FRANÇOIS MARTIN POUR LA RECHERCHE - DR

Un algorithme peut produire un résumé par extraction exploitable sans compréhension fine du texte

un livre dans une page, toute une page dans une phrase, et cette phrase dans un mot, c'est moi (1). »
À partir des années 1950, les bases scientifiques des solutions automatisées se sont échelonnées. Des avancées significatives ont été réalisées en matière de traitement automatique des langues (TAL) et de recherche de l'information (RI). En particulier, on peut noter les travaux fondateurs de l'informaticien allemand Hans Peter Luhn, en 1950, et ceux de la chercheuse britannique Karen Spärck Jones, en 1980, consacrés au résumé de textes scientifiques. Sans oublier, bien sûr, l'Américain Julian Kupiec, qui a mis au point dans les années 1990 des systèmes d'extraction d'information encore plus poussés.
Ces avancées ont consolidé le résumé statistique par extraction à partir des années 2000, puis ce que l'on a appelé les algorithmes neuronaux de résumé génératif, à partir de 2016. Le premier identifie – sans avoir besoin d'apprentissage – les phrases les plus saillantes d'un texte en utilisant des caractéristiques de la représentation des phrases : les mots, leurs co-occurrences et leurs probabilités ainsi que la structure du document, la position des phrases, etc. Puis il les extrait et les assemble afin de constituer un résumé exploitable. Quant aux algorithmes neuronaux (ou algorithmes d'apprentissage profond), ils utilisent des représentations complexes des mots et des phrases à l'aide de réseaux de neurones artificiels organisés en unités interconnectées réparties en couches. L'entraînement de ces réseaux d'intelligence artificielle nécessite de vastes corpus d'apprentissage, ce qui permet d'établir des correspondances entre les entrées (par exemple, les phrases d'un texte) et les sorties du réseau (par exemple, le contexte des mots ou la génération d'une phrase pertinente par rapport au contenu) (2).
Cependant, même les plus puissants algorithmes restent, à ce jour, incapables d'analyser et de comprendre un texte comme le font les humains. Une explication à cela : les différentes langues possèdent une structure syntaxique et sémantique étayée par des phrases, qui sont elles-mêmes constituées de mots et de structures linguistiques complexes ; la langue écrite contient également des redondances, voire des fautes (de grammaire, de syntaxe ou de contenu), qui rendent son apprentissage difficile.

TRANSPOSER LA COMPRÉHENSION DANS LE DOMAINE DU CALCULABLE

Est-il pour autant utile qu'un algorithme comprenne un texte en profondeur pour en produire un résumé par extraction exploitable ? La réponse est non. Il suffit en effet qu'il soit capable de repérer des zones informatives – les morceaux de texte contenant des objets linguistiques intéressants comme les verbes d'action, les noms propres et les entités nommées – pour extraire des informations pertinentes, les hiérarchiser et les organiser afin de générer des synthèses. L'important étant de trouver un équilibre entre les informations que l'on souhaite retenir et ce qu'il est possible d'extraire d'un texte. Or, pour automatiser ce processus dans un ordinateur, il faut transposer le problème de la compréhension et du résumé dans le domaine du calculable. Cela requiert de substituer les objets linguistiques par une représentation abstraite susceptible d'être comprise par des machines, tout en préservant l'information contenue dans le texte.
Dans le cas du résumé automatique, comment transposer le problème de la compréhension dans le domaine du calculable ? Pour ce faire, il faut établir une représentation adéquate du texte, qui peut être un modèle vectoriel. Dans un espace vectoriel, les m mots constituant le lexique du document sont plongés dans un espace des p phrases. Ce qui constitue une matrice de $[p \times m]$ dimensions. L'analyse linguistique permet de normaliser et de réduire les variations des mots (déclinaisons de verbes) ou d'éliminer les mots ou les symboles dits « creux » ou peu porteurs d'information (articles, conjonctions, ponctuation). Les mots rares ou au contraire trop fréquents sont traités statistiquement afin de pondérer leur importance de façon adéquate. Le texte devient ainsi

Comment évaluer la qualité d'un résumé ?

La question de l'appréciation de la qualité d'un résumé reste un problème non résolu, les chercheurs n'y répondant qu'avec des solutions partielles. On peut regrouper les méthodes d'évaluation selon qu'elles sont manuelles ou automatiques. Les premières emploient des évaluateurs humains, qui lisent et notent un résumé en fonction de critères préétablis : cohérence, grammaticalité, pertinence, etc. C'est une démarche objective, mais peu pratique et onéreuse. Les secondes font appel aux algorithmes qui évaluent la qualité du résumé automatique en utilisant des résumés de référence créés par des humains. Une autre solution approximative consiste aussi à ramener ce problème dans le domaine du calculable, en comptant des éléments dans le résumé automatique puis dans les résumés humains, afin d'établir des statistiques appropriées pour mesurer une proximité lexicale.
Sur quels éléments peut-on compter ? Des séquences des mots ou n-grammes. Si n = 1, on parle d'unigrammes (mots) ; si n = 2, il s'agit de bigrammes (paires de mots). Si on dispose de résumés créés par des humains, ceux-ci sont utilisés comme références. Prenons pour exemple un résumé de référence R = « J'ai une plume… elle est jolie ! La plume de ma tante » et deux résumés automatiques : r1 = « elle est jolie » et r2 = « la plume de ma tante ». R, r1 et r2 seront représentés par leurs n-grammes. Pour simplifier, nous utiliserons uniquement des bigrammes. Alors on a : R = {j'ai une, une plume, plume elle, elle est, est jolie, jolie la, la plume, plume de, de ma, ma tante}, avec un total de 10 bigrammes ; r1 = {elle est, est jolie} et r2 = {la plume, plume de, de ma, ma tante}. r1 partage 2/10 bigrammes avec R et r2 en partage 4/10, donc r2 possède une « meilleure qualité » ou proximité au résumé humain que r1. Bien sûr, les calculs dans la réalité sont plus complexes car ils font intervenir plusieurs statistiques, mais l'esprit reste le même. La méthode a été implémentée dans un algorithme appelé Rouge (1) et elle est très utilisée par la communauté scientifique. D'autres méthodes sans référence existent, l'idée étant de mesurer le contenu sémantique d'un résumé au moyen d'approximations (lexicales ou mixtes) vis-à-vis du document source (2).

(1) Ch.-Y. Lin, *Text Summarization Branches Out, ACL, 74,* 2004.

(2) A. Louis et A. Nenkova, *Computational Linguistics, 39,* 267, 2013.

un objet abstrait susceptible d'être traité par des opérations mathématiques et statistiques et des méthodes probabilistes. En outre, cette représentation abstraite permet de travailler sur plusieurs langues avec les mêmes algorithmes. En effet, ces méthodes ont l'avantage d'être assez indépendantes de la langue, car un prétraitement approprié y est appliqué. Ainsi, les caractéristiques linguistiques (déclinaison de verbes, genres, temps) sont normalisées, la suppression de mots creux est appliquée. Ensuite, l'occurrence des mots restants est transformée numériquement, chaque mot étant représenté par une quantité pondérée. L'information véhiculée par le texte est donc présente, mais transposée dans une représentation différente, non pas dans un espace textuel mais mathématique. Les approches de graphes sont fort utiles dans cette représentation, car les liens entre les lignes de la matrice (les phrases) au moyen des colonnes (les mots) peuvent être déduits ou calculés par co-occurrence. Même si l'ordre des mots peut être perdu – selon une représentation dite en « sac-de-mots » –, cette transformation présente beaucoup d'avantages. Bien que grossière du point de vue linguistique, elle est très efficace algorithmiquement, car elle préserve les caractéristiques importantes du lexique qui véhicule l'information du texte, telles la fréquence, les co-occurrences ou encore la rareté.

De nombreux logiciels de résumé automatique ont ainsi vu le jour sur ce modèle. Cependant, bien peu de ces solutions « toutes faites » sont optimisées pour traiter des textes aussi complexes que les études scientifiques ! Voilà pourquoi le Laboratoire d'informatique de l'université d'Avignon (LIA) a conclu en 2019 une collaboration avec SciencePOD, société d'édition spécialisée dans la communication scientifique. Notre objectif : rendre accessible le contenu d'ar-

POUR EN SAVOIR PLUS

J.-M. Torres-Moreno, *Automatic Text Summarization*, Wiley, 2014.

ticles scientifiques à de plus larges audiences sur de grandes quantités de documents traités en parallèle. Ce qui nécessitait, selon nous, la mise au point d'algorithmes capables de générer en à peine quelques centièmes de seconde ce qu'on appelle des résumés contextualisés.

PRODUIRE UN CONTEXTE POUR LES LECTEURS NON INITIÉS

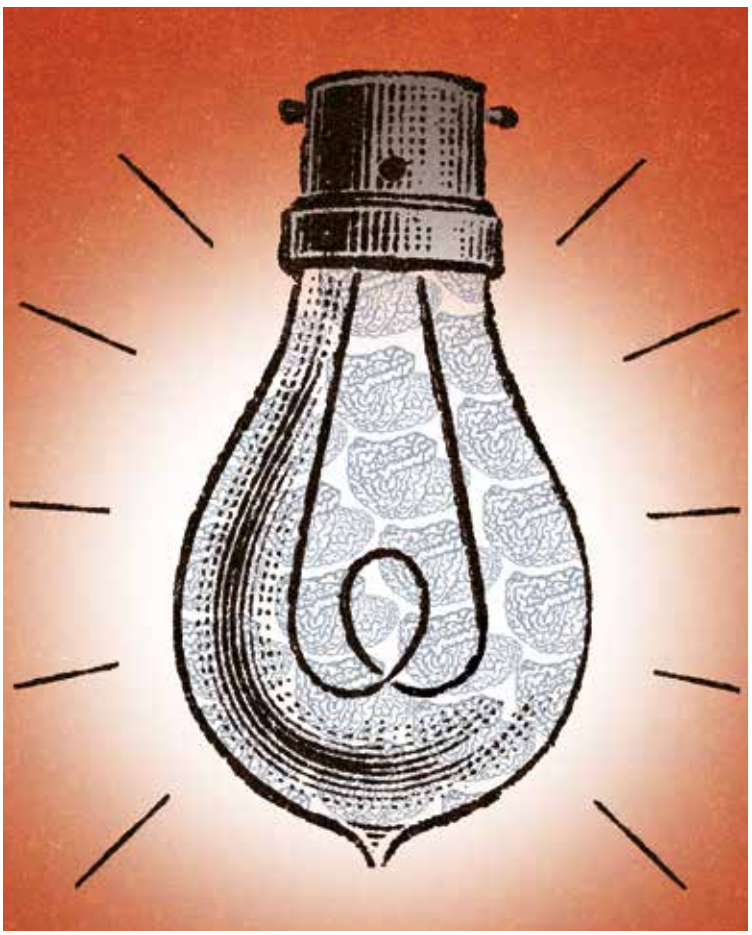
Pour obtenir ce type de résumé, il faut faire en sorte que les algorithmes extraient les phrases les plus importantes du document source de manière sélective et hiérarchisée. Pour ce faire, nous avons choisi d'utiliser des modèles à base de graphes : les n-grammes (lire l'encadré p. 119) représentent les sommets, et les arcs leurs liaisons statistiques (co-occurrence de mots, probabilités d'occurrence, pondération ou suppression statistique des mots, calcul d'entropie, etc.). Des techniques linguistiques permettent également de normaliser les mots (verbes à l'infinitif, noms au singulier, etc.) ou de calculer leur catégorie grammaticale (nom, verbe, adjectif, adverbe, ponctuation). Ensuite, nous avons conçu notre algorithme de telle manière qu'il produise un contexte pour les lecteurs qui ne seraient pas familiers du sujet de l'étude. En particulier, nous avons automatisé l'extraction de mots-clés, l'élucidation des acronymes et ajouté des définitions courtes des mots techniques.
En complément, nous avons créé un deuxième type d'algorithme pour extraire l'information clé de chaque étude. Nous avons ensuite fait en sorte que notre algorithme donne instantanément accès à cette information. En l'occurrence, en guidant l'algorithme à repérer certains termes, nous l'avons programmé pour qu'il extraie des méta-informations pertinentes sur l'auteur, son institution, etc. Nous avons centré la structure de ce type de résumé sur les réponses aux questions suivantes : quand et où l'étude a-t-elle été publiée ? Qui en sont les auteurs ? Où travaillent-ils ? Qu'ont-ils découvert ? Comment la recherche a-t-elle été conduite et quelles sont les pistes de futurs travaux ?
Une autre problématique à laquelle nous nous sommes intéressés a été de produire des résumés contextualisés d'articles issus de domaines scientifiques divers. Nous avons choisi de combiner plusieurs approches complémentaires afférentes à l'IA. Ces dernières incluent l'IA, le traitement automatique des langues, les méthodes d'analyse statistique, des méthodes de recherche d'information, l'analyse linguistique superficielle *(shallow parsing)* et la constitution et l'utilisation de ressources linguistiques structurées libres (ontologies, Wikipédia, thésaurus MeSH pour le domaine médical).
Plus ambitieux que le résumé extractif, le résumé dit « abstractif » s'appuie sur des modèles neuronaux d'apprentissage profond. Ces derniers produisent des nouvelles phrases – absentes du texte d'origine – censées être plus proches d'un résumé écrit par un humain. Mais l'inconvénient est que cette méthode requiert des don-

Cas pratique : résumons le présent article

Afin de tester l'efficacité de notre méthode, nous avons fait le résumé extractif de cet article. Voici le résultat : ***« Suivez l'histoire d'une collaboration combinant les connaissances en matière de résumé automatique de Juan-Manuel Torres-Moreno du Laboratoire informatique d'Avignon et le savoir-faire en matière de communication scientifique de Sabine Louët, fondatrice de la société d'édition numérique SciencePOD. Imaginez : le nombre d'études scientifiques publiées annuellement est passé de 972 000 en 1996 à 2,5 millions en 2018, selon des chiffres de la National Science Foundation, aux États - Unis ! Ces études ne possèdent pas forcément de résumé ou, quand il est présent, celui-ci contient souvent trop de termes techniques, inintelligibles à un public non expert du domaine. Rien de plus naturel alors que de se demander si l'intelligence artificielle ne pourrait pas être utile pour générer des résumés automatiques. »***

nées d'apprentissage massives (big data) et un temps d'apprentissage conséquent. Une autre limitation concerne la taille des documents d'entrée du système neuronal : des études récentes montrent qu'ils doivent se limiter à 2 000 mots (ou *tokens*) (3). Cela exclut un nombre important d'articles scientifiques sous peine de les tronquer afin que les algorithmes puissent le traiter. En outre, l'abstraction peut engendrer des phrases agrammaticales ou pas tout à fait justes par rapport au document source. Ainsi, selon notre expérience, l'apprentissage automatique n'est pas la meilleure stratégie pour faire du résumé d'études scientifiques. En revanche, en adoptant les résumés extractifs contextualisés, nous avons retenu un type de résumé qui cherche à garder les phrases saillantes. Après avoir extrait ces phrases-clés, nous les compressons davantage en éliminant des constituants non essentiels à la compréhension de la phrase après analyse du discours. Ensuite, nous assemblons ces phrases pour rendre un ordre logique au résumé sans se contenter de livrer des fragments de phrases-clés comme le font beaucoup de « résumeurs » génériques. Enfin, un contexte adéquat et les métadonnées complètent ce résumé.

De fait, l'approche extractive est plus simple à implémenter et plus rapide, car elle ne requiert pas d'apprentissage. Elle est aussi très qualitative, car elle est fondée sur des algorithmes statistiques et linguistiques classiques bien établis, améliorés grâce à notre savoir-faire. En outre, l'extraction se faisant sur des phrases complètes, les phrases sont grammaticalement correctes, même si cela ne garantit pas la cohérence du résumé final. Bien qu'ils soient encore loin des performances humaines, les résumés automatiques sont d'ores et déjà exploitables. Ils permettent au lecteur de décider ou non de lire le texte source dans son intégralité. Par ailleurs, le résumé automatique peut servir dans des systèmes informatisés de traitement d'information, tels les systèmes d'indexation documentaire, combinant résumés, termes et mots-clés pour permettre aux moteurs de recherche sur Internet de les repérer plus facilement. Enfin, ils peuvent aider les éditeurs scientifiques à présélectionner des études susceptibles d'être soumises à un comité de lecture. De grands éditeurs comme Elsevier, Springer Nature et Wiley ont ainsi lancé des expérimentations pour obtenir des résumés de leurs publications.



Nous envisageons aujourd'hui de raffiner les résumés en allant plus loin dans l'extraction d'information et dans les techniques d'IA. À l'heure où les publications scientifiques paraissent en accès libre *(open access)*, nous souhaitons proposer des résumés automatiques disponibles. Mais attention ! Ce n'est pas parce qu'un texte a fait l'objet d'un résumé et que sa teneur a été comprise que sa validité scientifique est assurée ! Les éditeurs dits « prédateurs » se multiplient en effet et bon nombre d'articles publiés, qui paraissent respectables de prime abord, ont un contenu scientifique limité. Le résumé automatique permet de gagner du temps dans la sélection d'études pertinentes, mais il ne se substitue pas au jugement de chacun. La sagacité humaine a encore de belles années devant elle ! ■

(1) Joseph Joubert, *Pensées, essais, maximes*, 1842 (sur gallica.bnf.fr).

(2) J. Zhang *et al.*, Pegasus: « Pre-training with Extracted Gap-sentences for Abstractive Summarization », « Proceedings of the 37th International Conference on Machine Learning », PMLR 119, 2020.

(3) A. Cohan *et al.*, « Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies », vol. 2, doi : 10.18653/v1/N18-2097, 2018.

JEAN-FRANÇOIS MARTIN POUR LA RECHERCHE