Moteur de recherche documentaire en langage naturel

Ying Zhang*, Matthieu Petit Guillaume*, Aurelien Krauth*

* Leviatan, 725 Boulevard Robert Barrier, 73100 Aix-les-Bains, France
y.zhang@leviatan.fr, matthieu@leviatan.fr, aurelien@leviatan.fr

1 Résumé étendu

Nous présentons un résumé étendu de l'article (Zhang et al., 2022) présenté à la journée scientifique LIFT-TAL 2022. À l'heure d'internet, il est de plus en plus facile et accessible de rechercher de l'information sur de nombreux types de sujets. Les archives documentaires et notamment celles générées par la presse spécialisée, jouent un rôle important chez les professionnels qui ont opéré leur transformation vers le numérique. Mais que deviennent ces archives documentaires et notamment les anciens numéros de magazines spécialisés ? Ceux-ci regorgent d'informations riches et précieuses dont la numérisation représente une solution efficace de stockage et un moyen rapide de recherche d'informations précises et pertinentes mis en oeuvre au travers d'une plateforme web. Dans le cadre de ce projet, nous avons stocké 1.1 To de magazines français initiaux aux formats pdf ou jpg.

Nous proposons un nouveau moteur de recherche documentaire en langage naturel, permettant d'accéder facilement à des informations précises dans des archives documentaires de masse. Le projet a été déployé dans un environnement de production avec un partenaire industriel et a été séparé en 4 composants principaux :

1. Prétraitement des magazines : Il s'agit d'un ensemble de prétraitements afin de transformer les magazines en version numérique. Nous avons principalement recours à un traitement OCR (Optical Character Recognition), au regroupement des textes par analyse de leurs informations géométriques, un ensemble de fonctions de nettoyage, une analyse linguistique et une transformation de paragraphe en word-embeddings (Yang et al., 2020).
2. Stockage des données : Les magazines originaux sont stockés dans un bucket AWS S3, les données numériques (sortie de l'étape 1) sont stockées dans un index Elasticsearch.
3. Filtrages des paragraphes à analyser pour une question posée : Nous avons 500 000 paragraphes stockés dans Elasticsearch. Étant donné qu'une question est posée par l'utilisateur, nous avons utilisé plusieurs stratégies de filtrages afin de récupérer uniquement les premiers 1000 paragraphes les plus pertinents pour des raisons de temps d'analyse.
4. Inférence de requête : Nous avons déployé une API de modèle MRC (Machine Reading Comprehension) afin de réaliser l'inférence de requête. Ce modèle a ensuite été ajusté, sur une base de modèle de langue CamemBERT (Martin et al., 2020) et de plusieurs jeux de données disponibles (Keraron et al., 2020; D'Hoffschmidt et al., 2020).

Le déploiement est basé sur un serveur GPU NVIDIA Tesla V100. Le temps de réponse du système est compris entre 3-5 secondes.

Nous avons testé ce système avec 4050 questions pré-annotées. 3703 questions ont des réponses et 348 questions liées aux bons documents mais n'ont pas de réponses précises. Nous proposons un maximum de 10 réponses pour chaque question selon l'ordre de score de fiabilité.

Dans les 3703 questions avec réponses, nous avons 3077 bonnes réponses retrouvées. Parmi ces 3077 bonnes réponses, 2906 réponses ont reçues un score de fiabilité du MRC élevé (>0.2), 171 réponses ont reçues un score de fiabilité faible (<0.2). Le système a proposé 355 mauvaises réponses mais tout de même trouvées dans le bon document. Enfin, le système a proposé 271 mauvaises réponses qui ne sont pas dans les bons documents.

Dans les 348 questions sans réponses, nous avons retrouvé 160 bons documents (79 questions ont reçues une réponse avec un score de fiabilité élevé et 81 questions ont reçues une réponse avec un score de fiabilité faible). 188 questions n'ont quant à elles pas pu être rattachées au bon document.

Notre plateforme web permet non seulement de présenter les résultats d'une requête, mais également de gérer un espace membre dédié afin que les utilisateurs puissent partager et contribuer aux contenus les plus pertinents associés aux sujets donnés.

Références

D'Hoffschmidt, M., W. Belblidia, Q. Heinrich, T. Brendlé, et M. Vidal (2020). FQuAD : French question answering dataset. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, Online, pp. 1193–1208.
English

Keraron, R., G. Lancrenon, M. Bras, F. Allary, G. Moyse, T. Scialom, E.-P. Soriano-Morales, et J. Staiano (2020). Project PIAF : Building a native French question-answering dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, pp. 5481–5490.

Martin, L., B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, et B. Sagot (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 7203–7219.

Yang, Y., D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernandez Abrego, S. Yuan, C. Tar, Y.-h. Sung, B. Strope, et R. Kurzweil (2020). Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, Online, pp. 87–94.

Zhang, Y., M. Petit Guillaume, et A. Krauth (2022). Documentary Research in Natural Language (D.R.N.L.) : Plateforme d'accès numérique aux archives documentaires en langage naturel. In *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, Marseille, pp. pp.74–83.