Utilisation des termes clés en recherche d'information - le cas du challenge TREC COVID

Adrian Chifu*, Bernard Dousset**,***
Josiane Mothe**,****

*LIS UMR CNRS 7020,
Aix-Marseille Université/Université de Toulon, France
adrian.chifu@univ-amu.fr,
https://adrianchifu.com
** Institut de Recherche en Informatique de Toulouse
UMR 5505 CNRS, Toulouse Prénom.Nom@irit.fr
http://www.irit.fr/∼Prénom.Nom
***Université Paul Sabatier, Université de Toulouse
****INSPE Université Toulouse Jean-Jaurès, Université de Toulouse

Résumé. En avril 2020, le NIST a proposé un premier cycle d'évaluation de recherche d'information comprenant un ensemble de sujets d'intérêt ou requêtes et une collection de documents scientifiques en lien avec la COVID-19. Cet article présente notre participation à ce challenge en considérant la première série de requêtes du challenge. Notre modèle s'appuie sur l'utilisation des termes clés plutôt que les usuels mots simples ou racines. Nous avons utilisé un modèle d'extraction de termes clés à base de n-grammes. Nous avons comparé les performances du moteur de recherche lorsque les documents sont représentés par des termes simples et par des termes clés. Nous montrons que certaines requêtes bénéficient de la représentation par des termes clés.

1 Introduction

Le NIST/TREC [1] a proposé un effort conjoint appelé TREC-COVID [2] au printemps 2020 (Voorhees et al., 2021). Comme les autres challenges proposés par TREC, TREC-COVID vise à rassembler des chercheurs du domaine de la recherche d'information pour évaluer les moteurs sur des tâches spécifiques. Ici, il s'agit de retrouver des documents scientifiques qui répondent à un sujet en lien avec la COVID-19. Cette campagne d'évaluation a lancé cinq cycles successifs d'évaluation. Chaque cycle propose des sujets d'intérêt pour la construction de requêtes et une collection de documents. Après la clôture d'un cycle, les jugements de pertinence correspondant aux documents pertinents de la collection pour chaque sujet d'intérêt sont également fournis.

Dans cet article, nous présentons les résultats officiels que nous avons obtenus lors de notre participation à cette campagne d'évaluation pour le premier cycle et les complétons par

1. https://trec.nist.gov
2. https://ir.nist.gov/covidSubmit/

des expérimentations supplémentaires ; nous menons une étude quantitative et qualitative de nos résultats.

Plus précisément, nous nous sommes intéressés à l'utilisation de termes clés ou groupes de mots. Concernant l'évaluation, nous nous appuyons sur les mesures quantitatives habituelles de recherche d'information. Nous menons ensuite une étude qualitative en nous intéressant à certains besoins spécifiques (requêtes) pour essayer de comprendre les cas où l'approche à base de termes clés fonctionne mieux qu'une approche à base de mots simples.

2 Tâche et collection

Le premier cycle d'évaluation proposé en avril 2020 dans le cadre de TREC-COVID comprend 30 sujets d'intérêt et une collection de plus de 45 000 publications scientifiques.Les données sont disponibles sur le site du NIST [3]. Cinq cycles successifs d'évaluation ont été proposés durant cette campagne d'évaluation, nous nous centrons dans cette étude sur le premier cycle : il s'agit du cycle auquel nous avons participé et pour lequel les jugements de pertinence pour les requêtes étaient disponibles pour réaliser les expérimentations supplémentaires (Section 4.2). D'un cycle d'évaluation à un autre, la collection de documents et le nombre de requêtes sont enrichis.

L'ensemble des documents retenus provient des données ouvertes "COVID-19 Open Research Dataset (CORD-19)" (Wang et al., 2020), publié le 10 Avril 2019 [4]. CORD-19 est un corpus d'articles scientifiques sur la COVID-19 et la recherche sur les coronavirus. Il est géré et maintenu par l'équipe de sémantique de l'Institut Allen afin de soutenir la recherche en traitement automatique des langues et la fouille de textes. Les documents sont les articles complets, contenant des informations comme le titre de l'article, les auteurs et leurs affiliations, les sections de l'article, le texte de l'article et les références bibliographiques.

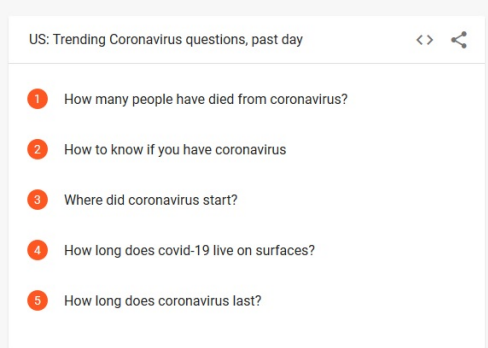
Trente sujets d'intérêt ont été retenus par les organisateurs ; ils correspondent aux questions que les citoyens se posaient au moment du début de la pandémie (Dousset et Mothe, 2020). Par exemple, les requêtes les plus populaires selon Google Trends le 27 avril 2020 portaient sur le nombre de morts, les symptômes, l'origine, la survie du virus sur les surfaces (cf. Figure 1a). Les sujets choisis par TREC sont, par exemple, l'origine du Coronavirus, sa survie à l'extérieur ou les premiers symptômes (cf. Figure 1b).

Pour chaque sujet d'intérêt, la liste des documents pertinents est fournie après la clôture du cycle correspondant.

3 Représentation et recherche des documents

Les approches que nous avons proposées pour traiter cette tâche de recherche d'information relèvent d'une approche usuelle de comparaison des requêtes avec les documents. L'originalité relève (a) de la représentation à base de termes clés (b) l'approche d'appariement.

3. https://ir.nist.gov/covidSubmit/data.html
4. https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/historical_releases.html
5. https://trends.google.com/trends/story/US_cu_4Rjdh3ABAABMHM_en, consulté le 27/04/2020.



(a) *Google Trends*[5] *- Comment Coronavirus est recherché.*

Query: coronavirus origin
Related question: What is the origin of COVID-19

Query: Coronavirus outside body
Related question: how long can the coronavirus live outside the body

Query: Coronavirus early symptoms
Related question: what are the initial symptoms of Covid-19?

(b) *Exemple de topics TREC.*

FIG. 1. *Les requêtes populaires de Google Trends liées au Coronavirus et des exemples de topics TREC.*

3.1 Structure des documents

Les documents de la collection étant des articles scientifiques complets, nous avons supprimé certaines sections comme les références bibliographiques qui auraient pu faire dévier la représentation des documents de leur contenu propre. En effet, les références, même s'ils sont liés au contenu de l'article qui les utilisent, ne sont généralement pas pertinentes pour représenter le document lui-même. Afin d'indexer le corpus, nous avons donc concaténé, pour chaque document, le titre, les titres des sections et le texte de l'article.

3.2 Structure des besoins d'information

Les trente besoins en information du premier cycle d'évaluation de la tâche TREC-COVID ont la structure suivante :

— une balise `<query>` contenant les mots de la requête (par exemple : `<query>coronavirus under reporting </query>`);
— une balise `<question>` contenant une question qui représente le besoin d'information (par exemple : `<question>how has lack of testing availability led to underreporting of true incidence of Covid-19?</question>`)
— une balise `<narrative>` contenant des détails sur ce qui caractérise un document pertinent pour le besoin d'information (par exemple : `<narrative>Looking for studies answering questions of impact of lack of complete testing for Covid-19 of incidence and prevalence of Covid-19.</narrative>`).

Dans nos recherches, nous exploitons les contenus des balises `<query>` pour construire les requêtes.

3.3 Unité d'indexation

Nous avons ensuite extrait les termes représentatifs des textes selon deux niveaux différents : les radicaux par Indri (version 5.11) [6] et les termes clés.

L'indexation par les radicaux s'appuie sur Indri, en supprimant les mots vides et en appliquant le raciniseur de Krovetz (1993). Ce raciniseur utilise l'analyse flexionnelle et un dictionnaire pour déterminer les formes correctes des mots avant de supprimer les terminaisons des mots. (Schofield et Mimno, 2016) ont montré qu'il s'agissait d'un raciniseur parmi les plus efficaces et il est implanté par défaut sur le moteur Indri (version 5.11) [7] que nous utilisons par ailleurs pour l'indexation et la recherche. Indri est un moteur de recherche basé sur les modèles de langue, très utilisé dans la littérature et qui produit des résultats compétitifs. C'est la raison pour laquelle nous l'avons choisi.

L'indexation par les termes clés s'appuie sur Tétralogie [8]. Il s'agit d'un outil de veille qui permet l'analyse et de fouille de documents textuels. Son développement a débuté il y a plusieurs décennies et il a été utilisé dans de nombreuses actions de veille scientifiques pour des industriels et des laboratoires (Gay et Dousset, 2005). Il intègre une fonction d'extraction de termes clés basée sur la détection de n-grammes de mots.

Dans Tétralogie, la détection considère d'abord le lexique des termes contenant un tiret ("-") comme base. L'outil d'extraction recherche ensuite les n-grammes fréquents tout en prenant en compte les mots outils de la langue soit en les conservant (exemple "de" est gardé), soit en les supprimant (exemple "et" est supprimé). Un seuil sur la fréquence du n-gramme dans différents documents permet de le sélectionner ou au contraire de l'écarter. Le document est ensuite ré-écrit complètement en remplaçant les occurrences des mots simples par les termes clés extraits.

Le document ré-écrit comprend ainsi des mots simples (lorsque aucun n-gramme n'a été extrait) et des termes clés (les n-grammes). L'indexation des documents par des termes clés est réalisée comme dans le cas des radicaux, mais plutôt que le texte original, le document ré-écrit est indexé par Indri.

3.4 Appariement requête/documents

Nous nous sommes intéressés à deux types d'appariement :

— Appariement simple dans lequel les documents doivent contenir les termes de la requête pour être retrouvés. C'est cette stratégie qui a été utilisée dans notre participation officielle à TREC COVID, selon deux variantes précisées en section 4.1.

— Indri BM25 : nous nous appuyons sur le modèle BM25 Robertson et Jones (1976) avec 1000 documents retrouvés par requête et en considérant les paramètres par défaut ($k_1 = 1.2$, $k_3 = 7$, $b = 0.75$). Ce type d'appariement a été utilisé dans les expériences complémentaires présentées en section 4.2.

6. https://www.lemurproject.org/indri.php
7. https://www.lemurproject.org/indri.php
8. https://atlas.irit.fr/PIE/Outils/Tetralogie.html

Les méthodes d'appariement combinées avec les types d'indexation nous ont permis d'obtenir différentes listes de documents retrouvés que nous avons évaluées grâce à l'outil nommé `trec_eval` [9].

4 Résultats et discussions

4.1 Participation officielle

Lors de notre participation à la tâche, nous avons choisi deux stratégies de recherche qui s'appuient sur l'utilisation des termes clés avec un appariement simple (sans utiliser Indri donc).

— Stratégie_1 : les documents sont ré-écrits avec les termes clés. Les requêtes (les balises `query` des besoins d'information) sont également ré-écrites avec les termes clés existant dans les documents. Il suffit ensuite qu'un des termes de la requête apparaisse dans la requête pour qu'il soit retrouvé. Les documents ne sont pas ordonnés, ils sont restitués dans l'ordre de leur apparition. Cette stratégie favorise le rappel que la mesure AP prend en compte pour son calcul.

— Stratégie_2 : elle se distingue de la stratégie précédente au moment de l'appariement : les documents sont ordonnés en fonction du nombre de termes de la requête qu'ils contiennent. Cette stratégie est donc davantage orientée vers la précision.

Si l'on ne considère que la performance moyenne sur un ensemble de requête, ces deux stratégies n'ont pas été très performantes en comparaison avec une approche par mots simples dans un modèle BM25. La Table 1 compare les résultats selon différentes mesures. Nous avons retenu :

— la MAP (mean average precision), moyenne des AP (average precision) de l'ensemble des requêtes. La mesure AP est définie comme suit :

$$AP = \frac{\sum_{k=1}^{n} (P@k \times rel@k)}{\text{nombre total de documents pertinents}},$$

où k est le rang d'un document de la liste retrouvée, n est le nombre total de documents retrouvés, $P@k$ est la précision à k documents et $rel@k$ est une fonction égale à 1 si le document de rang k est un document pertinent, et à 0 sinon. Cette mesure porte sur tous les documents pertinents et les documents pertinents non retrouvés obtiennent un score de précision de zéro. La mesure AP suppose que l'utilisateur souhaite trouver un grand nombre de documents pertinents pour chaque requête, ainsi, l'amélioration du rappel a un impact sur cette mesure (Stratégie_1) ;

— la P@5 (précision à 5 documents) moyenne sur l'ensemble des requêtes des précisions obtenues pour 5 documents retrouvés (Stratégie_2) ;

— le ndcg@10 (Normalized Discounted Cumulative Gain à 10 documents), une mesure de pertinence graduelle normalisée, avec un seuil fixé à dix documents qui pénalise les documents très pertinents apparaissant plus bas dans une liste de résultats de recherche. En effet, la valeur de pertinence est réduite de manière logarithmique proportionnellement à la position du résultat.

9. `https://trec.nist.gov/trec_eval/` : outil permettant d'évaluer les résultats d'un moteur de recherche d'information selon une variété de mesures.

TAB. 1. *Résultats officiels TREC COVID comparés à BM25.*

Stratégie	MAP	P@5	ndcg@10
BM25	0,182	0,493	0,411
Stratégie_1	0,012	0,113	0,082
Stratégie_2	0,022	0,207	0,150

En moyenne sur les trente requêtes, BM25 sur les radicaux est une meilleure stratégie que nos deux stratégies. En moyenne également, parmi nos deux stratégies officielles, la Stratégie_2 est meilleure que la Stratégie_1.

Nous avons également analysé les résultats par requête. Nous nous sommes intéressés à la P@5. Quelque soit la requête considérée, la Stratégie_2 est meilleure que la Stratégie_1. En revanche, nous avons pu observer que pour les requêtes 8 et 18, la Stratégie_2 est supérieure à BM25. Dans le cas de la requête 8 "coronavirus under reporting", l'extraction du terme clé "UNDER REPORTING" s'est avérée cruciale avec une P@5 de 0 pour BM25 et de 0,4 avec la Stratégie_2. Nous avons donc poursuivi les expérimentations sur ce jeu de données et l'utilisation de termes clés, hors compétition.

4.2 Autres expérimentations

Dans les autres expérimentations, nous avons souhaité combiner les approches BM25 avec Indri et l'extraction des termes clés.

Nous avons donc indexé via Indri les documents ré-écrits avec les termes clés et nous sommes ensuite appuyés sur BM25 pour la recherche. Cette stratégie est nommée MTMBM25. La requête (partie *query*) est également ré-écrite avec les termes clés. La Table 2 présente quelques résultats.

TAB. 2. *Résultats non-officiels comparés à BM25.*

Stratégie	MAP	P@5	ndcg@10
BM25	0,182	0,493	0,411
MTMBM25	**0,183**	**0,573**	**0,425**

Nous pouvons noter que cette nouvelle stratégie, MTMBM25, s'avère efficace quelque soit la mesure en moyenne sur les 30 requêtes. D'autres comparaisons pourraient être ajoutées, comme par exemple BM25 sur les mots sans racinisation, sans la suppression des mots vides ou en utilisant d'autres raciniseurs.

4.3 Analyse détaillée

Nous nous sommes ensuite intéressés aux requêtes individuelles.

Quelque soit la mesure considérée, la majorité des requêtes bénéficient du traitement par termes clés (cf. Table 4 qui montre le nombre de requêtes pour lesquelles une stratégie est meilleure que les autres pour les trois mesures d'évaluation considérées). Cette stratégie est particulièrement efficace pour les hautes précisions. Pour la P@5, seules trois requêtes devrait

être traitées par BM25 avec les mots simples plutôt qu'avec les termes clés pour obtenir de meilleurs résultats.

La Table 3 présente le détail des résultats pour BM25 et MTMBM25, par mesure de performance et par requête.

TAB. 3. *Le détail par mesure de performance, par exécution et par requête. Les meilleures valeurs sont écrites en gras (en cas d'égalité, aucune des deux valeurs n'est marquée en gras).*

	AP/MAP		P@5		ndcg@10	
Requêtes	BM25	MTMBM25	BM25	MTMBM25	BM25	MTMBM25
1	**0,129**	0,121	0,400	0,400	0,307	**0,372**
2	0,150	**0,173**	0,600	0,600	**0,507**	0,442
3	0,039	**0,040**	**0,200**	0,000	**0,082**	0,076
4	0,001	**0,017**	0,000	0,000	0,000	0,000
5	0,021	**0,038**	0,400	**0,600**	0,220	**0,425**
6	0,162	**0,170**	0,600	**0,800**	0,477	**0,494**
7	0,089	**0,102**	0,200	0,200	**0,225**	0,085
8	0,008	**0,009**	0,000	0,000	**0,074**	0,000
9	0,122	**0,143**	0,400	**0,600**	0,292	**0,368**
10	**0,456**	0,434	1,000	1,000	0,849	**0,878**
11	**0,056**	0,044	0,200	0,200	**0,220**	0,126
12	0,146	**0,163**	0,800	**1,000**	0,627	**0,705**
13	0,032	**0,097**	0,400	**0,600**	0,200	**0,231**
14	**0,316**	0,261	0,800	0,800	0,523	**0,549**
15	0,032	**0,052**	0,400	0,400	0,117	**0,204**
16	0,162	**0,174**	0,400	0,400	**0,317**	0,263
17	**0,257**	0,235	0,600	**1,000**	0,754	**0,852**
18	0,233	**0,258**	0,400	**0,800**	0,506	**0,550**
19	0,125	**0,135**	0,600	0,600	**0,514**	0,439
20	0,151	**0,179**	0,400	**0,800**	0,337	**0,589**
21	0,109	**0,116**	0,400	**0,600**	0,362	**0,413**
22	**0,178**	0,163	**1,000**	0,800	**0,658**	0,573
23	0,290	**0,305**	0,600	0,600	**0,537**	0,528
24	**0,284**	0,248	**0,800**	0,600	**0,548**	0,420
25	**0,101**	0,079	0,400	0,400	**0,246**	0,195
26	0,122	**0,124**	0,400	0,400	**0,290**	0,249
27	0,192	**0,194**	0,400	**0,600**	0,497	**0,572**
28	0,590	**0,602**	0,600	0,600	**0,820**	0,766
29	**0,247**	0,169	0,400	**0,800**	0,445	**0,587**
30	0,645	**0,650**	1,000	1,000	0,791	0,791
Moyenne	0,182	**0,183**	0,493	**0,573**	0,411	**0,425**

Lorsque l'on considère la P@5, la différence est particulièrement intéressante pour 4 requêtes (différence d'au moins 0.4 en faveur de la stratégie par termes clés). Les termes clés

TAB. 4. *Pour les hautes précisions (P@5), l'utilisation des termes clés améliore 11 requêtes et n'en pénalise que 3 sur les 30. Pour les autres mesures l'utilisation des termes clés favorise beaucoup plus de requêtes qu'elle n'en pénalise.*

	BM25	**MTMBM25**	**Egalité**
AP	9	21	0
P@5	3	11	16
ndcg@10	13	15	2

"ON SURFACE", "CLINICAL TRIAL", "ACE INHIBITOR" et "DRUG REPURPOSING" font vraiment la différence. D'autres termes clés sont importants comme "ANIMAL MODEL".

En ce qui concerne la mesure AP, même si en moyenne (MAP), les résultats sont très proches (0,182 pour BM25 et 0,183 pour MTMBM25), nous pouvons constater que 70% des requêtes obtiennent des performances supérieures avec MTMBM25. Les améliorations les plus importantes (différence directe) ont lieu pour les requêtes 13 et 20, respectivement. Dans le cas de la requête 20 "coronavirus and ACE inhibitors", c'est l'extraction du terme clé "ACE INHIBITOR" qui fait la différence.

Lorsque l'on considère la mesure ndcg@10, nous pouvons comptabiliser 15 requêtes améliorées par MTMBM25 et 2 requêtes avec des performances identiques entre les exécutions BM25 et MTMBM25. Dans le cas de cette mesure, les valeurs en moyenne sont très proches (0,411 pour BM25 et 0,425 pour MTMBM25), comme pour la MAP. La meilleure amélioration est pour la requête 20 (0.251 de différence). La différence suivante (0.205 de différence) a lieu pour la requête 5. Cette requête non-transformée est : "animal models of COVID-19" ; l'extraction du terme clé "ANIMAL MODEL" est ici décisive.

Pour certaines requêtes, il n'existe pas de différence entre la requête initiale et la requête réécrite. Donc, l'impact sur la différence de performance provient de la ré-écriture des documents indexés et des pondérations des index résultantes.

5 Etat de l'art

5.1 Tâche COVID

En se basant sur les soumissions effectuées par des participants à la tâche TREC-COVID, Chen et Hersh (2020) ont proposé et évalué une taxonomie de caractéristiques des systèmes de RI associés à une haute performance dans le cadre TREC-COVID. L'analyse multivariée qu'ils ont réalisée montre que la façon de définir les requêtes joue un rôle important dans l'amélioration des performances. Ce résultat est en accord avec nos résultats : lorsque la requête contient des termes clés composés, nous avons pu voir l'efficacité de les considérer comme tels. Un autre facteur important est le retour de pertinence (*pseudo-relevance feedback*) qui est utile pour améliorer la performance ; nous n'avons pas encore étudié ce facteur. En revanche, l'analyse proposée ne permet pas de définir de façon générale quels sont les modèles et stratégies qui sont liés à une haute performance sur la tâche TREC-COVID.

A notre connaissance, peu d'articles scientifiques ont été publiés suite à la participation à la tâche TREC-COVID. Une liste est disponible et sera mise à jour au fur et à mesure des publications [10]

Lima et al. (2020) ont combiné des méthodes standards, des méthodes de l'état de l'art et des heuristiques de méta-recherche. Ils ont conclu que les jugements de pertinence ne peuvent pas être estimés de manière utile par des non-experts, que la source de la littérature scientifique est un indicateur fort de la pertinence, que les plongements de mots ne sont pas plus efficaces par rapport aux représentations basées sur les fréquences et que le volume de données est trop faible pour appliquer des méthodes avancées d'apprentissage automatique. Notre approche peut donc être vue comme un complément ; nous avons montré que la prise en compte des termes clés est efficace.

Wang et al. (2020) ont proposé une approché basée sur le modèle d'apprentissage actif continu (CAL) et ses variantes. Leur système a été classé parmi les premiers de la tâche. Parmi d'autres méthodes proposées dans des participations bien classées, nous pouvons mentionner (i) un classement initial avec et un reclassement avec SciBERT (un modèle neuronal de reranking pré-entraîné sur des textes scientifiques) entrainé sur MS-MARCO (une grande collection de données contenant des paragraphes de questions et de réponses, en langage naturel et de domaine générique), en utilisant le titre et le résumé des articles, ou (ii) LambdaRank avec 8 caractéristiques dérivées de 4 modèles de pertinence non supervisés, en exploitant également des entités nommées.

A notre connaissance, aucun autre participant à la tâche n'a utilisé les termes clés dans la tâche TREC-COVID ; il s'agit cependant d'une approche qui a déjà été utilisée en RI dans d'autres cadres. Par exemple, Datta et al. (2017) ont proposé un modèle multimodal d'extraction de termes clés basé sur un graphe qui saisit la relation entre les mots, en termes d'information réciproque et de retour de pertinence. Ils appliquent la méthode Fisher-LDA pour déterminer les pondérations appropriées pour chaque modalité. Song et al. (2005) ont introduit une technique de requêtage automatique pour identifier l'ensemble des documents afin d'extraire des relations prédéfinies à partir d'un texte. Leur modèle est conçu pour récupérer des documents correspondant plus précisément à la requête initiale en étendant les requêtes à chaque fois que le processus d'extraction de termes clés se répète dans une base de données. Ils utilisent l'extraction de termes clés en conjonction avec des ontologies de référence pour l'expansion de requêtes, dans le domaine biomédical.

5.2 Extraction de termes clés

De nombreuses méthodes ont été proposées dans la littérature pour l'extraction de termes clés de façon non supervisée comme cela est le cas dans Tétralogie. Ces méthodes s'appuient souvent sur des considérations statistiques comme dans Ding et al. (2011) avec la mesure TF-IDF (Sparck Jones, 1972) ou la mesure *C-values* dans Lossio-Ventura et al. (2013) et considèrent le problème comme celui d'un ordonnancement, les termes candidats les mieux classés sont conservés.

Tomokiyo and Hurst utilisent deux scores complémentaires : le *niveau* pour lequel une séquence de mots peut être considérée comme un terme clés et l'*informativité* qui considère les termes clés qui illustrent le mieux l'idée principale du document (Tomokiyo et Hurst, 2003).

10. `https://ir.nist.gov/covidSubmit/bib.html`

Dans les méthodes à base de graphe, le principe est de construire un graphe de mots ou de groupes de mots (les nœuds sont les termes clés candidats et une arête relie deux nœuds si les termes clés candidats sont liées). Les arêtes et leur poids peuvent être calculés en utilisant des cooccurrences (Matsuo et Ishizuka, 2004; Mihalcea et Tarau, 2004; Wan et Xiao, 2008) ou des liens sémantiques Grineva et al. (2009) par exemples. Les nœuds sont ensuite classés en fonction des propriétés du graphe, telles que la centralité (Lahiri et al., 2014), le degré, etc. Ce type d'approche est considérée comme l'une des meilleures solutions pour l'extraction non supervisée de termes clés (Beliga, 2014). Par exemple, Boudin a comparé différentes mesures de centralité. Lahiri étend l'approche de Boudin en comparant d'autres mesures de centralité et en utilisant davantage de corpus (Lahiri et al., 2014) et conclut, comme Boudin, que la mesure de centralité la plus simple surpasse les autres. Dans ces deux travaux, deux nœuds sont connectés s'ils cooccurrent dans une fenêtre de T mots mais les autres types de relations sémantiques entre les nœuds ne sont pas considérés.

Certaines méthodes ont introduit l'utilisation d'informations sémantiques. Dans SemanticRank, deux nœuds sont liés s'ils sont sémantiquement apparentés dans WordNet ou le thésaurus Wikipedia (Tsatsaronis et al., 2010). TopicRank (Bougouin et al., 2013) regroupe les mots en grappes et les utilise ensuite comme nœuds de graphe pour l'extraction de phrases clés. Cette méthode évite la redondance et considère que les phrases-clés qui partagent des mots appartiennent au même sujet. TopicalPagerank (Liu et al., 2010) et SingleTopicalPageRank (Sterckx et al., 2015) utilisent Latent Dirichilet Alllocation (Blei et al., 2003) pour favoriser le classement des mots en fonction d'un sujet spécifique.

Enfin, des méthodes utilisent des encapsulations de mots (word embeddings) (Wang et al., 2014; Mothe et al., 2018) pour créer un graphe de mots, puis utilisent des méthodes de l'état de l'art comme l'algorithme PageRank pondéré pour calculer le score de chaque nœud et les classer.

6 Conclusion

Dans cet article, nous avons présenté la tâche TREC COVID qui a été initiée au début de la pandémie en Avril 2020 et qui a pour objet de proposer des méthodes pour retrouver des articles scientifiques sur un sujet d'intérêt. Nous avons également présenté notre participation à cette tâche sur la première série de sujets d'intérêt.

Nous avons présenté des expérimentations complémentaires faisant appel à l'utilisation de termes clés. Les termes clés nous paraissent plus porteurs d'information que les termes simples et les sujets autour du Coronavirus nous ont semblé adaptés pour tester cette hypothèse.

Nos résultats ont montré l'apport des termes clés pour un certain nombre de requêtes et l'amélioration globale des résultats lors de l'utilisation des termes clés plutôt que des termes simples.

Dans les travaux futurs, nous aimerions nous appuyer sur des travaux récents d'extraction automatique de termes clés afin d'affiner nos résultats, les algorithmes Yake ! (Campos et al., 2018) et celui que nous avons proposé dans (Mothe et al., 2018) ont particulièrement retenu notre attention. YAKE ! est un autre algorithme d'extraction automatique de mots-clés, non supervisé, qui repose sur des caractéristiques statistiques extraites de documents individuels pour sélectionner les mots-clés les plus pertinents d'un texte. Une autre approche complémentaire sera d'utiliser le retour de pertinence automatique.

7 Remerciements

Ce travail est partiellement soutenu par le projet PREVISION, qui a reçu un financement du programme de recherche et d'innovation Horizon 2020 de l'Union européenne sous le n° 833115 (`https://cordis.europa.eu/project/id/833115`). Ce document reflète le point de vue des auteurs et la Commission n'est pas responsable de l'usage qui pourrait être fait des informations qu'il contient.

Références

Beliga, S. (2014). Keyword extraction : a review of methods and approaches. *University of Rijeka, Department of Informatics, Rijeka*.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research 3*(Jan), 993–1022.

Bougouin, A., F. Boudin, et B. Daille (2013). Topicrank : Graph-based topic ranking for keyphrase extraction. In *Int. Joint Conf. on NLP*, pp. 543–551.

Campos, R., V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, et A. Jatowt (2018). Yake ! collection-independent automatic keyword extractor. In *European Conference on Information Retrieval*, pp. 806–810. Springer.

Chen, J. et W. R. Hersh (2020). A comparative analysis of system features used in the trec-covid information retrieval challenge. *medRxiv*.

Datta, D., S. Varma, R. Chowdary C., et S. K. Singh (2017). Multimodal retrieval using mutual information based textual query reformulation. *Expert Systems with Applications 68*, 81–92.

Ding, Z., Q. Zhang, et X. Huang (2011). Keyphrase extraction from online news using binary integer programming. In *IJCNLP*, pp. 165–173.

Dousset, B. et J. Mothe (2020). Getting insights from a large corpus of scientific papers on specialisted comprehensive topics–the case of covid-19. *arXiv preprint arXiv :2005.00485*.

Gay, B. et B. Dousset (2005). Les réseaux d'alliances stratégiques dans le domaine des anticorps monoclonaux : étude longitudinale. In *Journées sur les systèmes d'information élaborée*.

Grineva, M., M. Grinev, et D. Lizorkin (2009). Extracting key terms from noisy and multi-theme documents. In *Int. conf. on WWW*, pp. 661–670.

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'93*, pp. 191–202.

Lahiri, S., S. R. Choudhury, et C. Caragea (2014). Keyword and keyphrase extraction using centrality measures on collocation networks. *Preprint arXiv :1401.6571*.

Lima, L., C. Hansen, C. Hansen, D. Wang, M. Maistro, B. Larsen, J. Simonsen, et C. Lioma (2020). Denmark's participation in the search engine trec covid-19 challenge : Lessons learned about searching for precise biomedical scientific information on covid-19.

Liu, Z., W. Huang, Y. Zheng, et M. Sun (2010). Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pp. 366–376. Association for Computational Linguistics.

Lossio-Ventura, J. A., C. Jonquet, M. Roche, et M. Teisseire (2013). Combining c-value and keyword extraction methods for biomedical terms extraction. In *LBM : Languages in Biology and Medicine*.

Matsuo, Y. et M. Ishizuka (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools 13*(01), 157–169.

Mihalcea, R. et P. Tarau (2004). Textrank : Bringing order into texts. ACL.

Mothe, J., F. Ramiandrisoa, et M. Rasolomanana (2018). Automatic keyphrase extraction using graph-based methods. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pp. 728–730.

Robertson, S. E. et K. S. Jones (1976). Relevance weighting of search terms. *Journal of the American Society for Information science 27*(3), 129–146.

Schofield, A. et D. Mimno (2016). Comparing apples to apple : The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics 4*, 287–300.

Song, M., I.-Y. Song, et K. J. Lee (2005). Automatic unsupervised keyphrase-based query expansion for biomedical domain. In *Knowledge Management : Nurturing Culture, Innovation, and Technology*, pp. 209–220. World Scientific.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation 28*(1), 11–21.

Sterckx, L., T. Demeester, J. Deleu, et C. Develder (2015). Topical word importance for fast keyphrase extraction. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 121–122. ACM.

Tomokiyo, T. et M. Hurst (2003). A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions : analysis, acquisition and treatment-Volume 18*, pp. 33–40. Association for Computational Linguistics.

Tsatsaronis, G., I. Varlamis, et K. Nørvåg (2010). Semanticrank : ranking keywords and sentences using semantic graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 1074–1082. Association for Computational Linguistics.

Voorhees, E., T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, et L. L. Wang (2021). Trec-covid : constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, Volume 54, pp. 1–12. ACM New York, NY, USA.

Wan, X. et J. Xiao (2008). Single document keyphrase extraction using neighborhood knowledge. In *AAAI*, Volume 8, pp. 855–860.

Wang, L. L., K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. Raymond, D. S. Weld, O. Etzioni, et S. Kohlmeier (2020). Cord-19 : The covid-19 open research dataset.

Wang, R., W. Liu, et C. McDonald (2014). Corpus-independent generic keyphrase extraction using word embedding vectors. In *Software Engineering Research Conference*, pp. 39.

Wang, X. J., M. R. Grossman, et S. G. Hyun (2020). Participation in TREC 2020 COVID Track Using Continuous Active Learning. *arXiv e-prints*, arXiv :2011.01453.

Summary

In April 2020, NIST proposed a first round of information retrieval evaluation including a set of topics of interest or queries and a collection of scientific documents related to COVID-19. This paper presents our participation to this challenge by considering the first series of queries. Our model relies on the use of key phrases rather than the usual single words or roots. We used an n-gram based key phrase extraction model. We compared the performance of the search engine when documents are represented by single terms and by key phrases. We show that some queries benefit from the key term representation.