Enrichissement sémantique non supervisé de longs documents spécialisés pour la recherche d'information

Oussama Ayoub

HESAM UNIVERSITÉ

le cnam

ÉCOLE DOCTORALE Sciences des Métiers de l'Ingénieur
Cédric & DVRC

THÈSE

présentée par : **Oussama AYOUB**
soutenue le : **22 décembre 2023**

pour obtenir le grade de : **Docteur d'HESAM Université**

préparée au : **Conservatoire national des arts et métiers**

Discipline : **CNU 27**

Spécialité : **Informatique**

Enrichissement sémantique non supervisé de longs documents spécialisés pour la recherche d'information

THÈSE dirigée par :
M. Nicolas TRAVERS Professeur, DVRC, Devinci Higher Education, France

et co-encadrée par :
M. Christophe RODRIGUES Enseignant-Chercheur, DVRC, Devinci Higher Education, France

Jury

Mme Salima BENBERNOU	Professeure des Universités, LIPADE, Université Paris-Cité	Rapportrice
Mme Nedra MELLOULI	Professeure des Universités, LIASD, Université Paris 8	Présidente Rapportrice
M. Aomar OSMANI	Maître de Conférences HDR, LIPN, Université Sorbonne Paris Nord	Examinateur
Mme Haïfa ZARGAYOUNA	Maître de Conférences, LIPN, Université Sorbonne Paris Nord	Examinatrice

THÈSE

Remerciements

Je tiens avant tout à exprimer ma profonde et sincère gratitude à mon directeur de thèse, Nicolas Travers, pour son accompagnement exceptionnel, son soutien constant et sa confiance en moi tout au long de ce voyage académique. Au-delà de son rôle de directeur de thèse, je tiens à le remercier pour son écoute, ses conseils avisés et son soutien moral durant les moments de doute et de challenge. Sa présence rassurante et son optimisme ont été des piliers sur lesquels je me suis souvent appuyé.

Je suis profondément reconnaissant à mon co-encradrant de thèse, Christophe Rodrigues. Dès les premiers jours de mon projet de recherche, son expertise profonde et sa passion pour le sujet ont été une source d'inspiration inépuisable. Sa vision éclairée et son approche pédagogique m'ont permis de surmonter les obstacles inhérents à la recherche scientifique et de m'ouvrir à de nouvelles perspectives de réflexion. Nos discussions, ses conseils avisés et sa bienveillance ont façonné cette expérience doctorale en une aventure intellectuelle et humaine des plus enrichissantes.

Un grand merci à Ghislain d'Aramon, Directeur général de Seville More Helory, qui a été présent depuis le début de ma carrière professionnelle et m'a offert les différentes opportunités qui m'ont permis d'arriver à l'obtention de ce diplôme. Un long chemin a été parcouru durant nos années de collaboration et je te remercie de m'avoir accompagné jusqu'au bout.

Je souhaite aussi adresser ma gratitude à Aomar Osmani et Raphaël Fournier-S'niehotta, les deux membres du comité de suivi de cette thèse. Votre capacité à identifier les points forts ainsi que les axes d'amélioration de mon travail a été d'une aide inestimable. Merci pour votre engagement à examiner mon travail et pour le temps consacré à la préparation de vos retours durant ces comités de suivis. Vos commentaires constructifs et vos encouragements m'ont motivé à persévérer tout au long de mes travaux de recherche.

Je remercie Salima BENBERNOU, Nedra MELLOULI, Aomar OSMANI et Haïfa ZARGAYOUNA,

d'avoir accepté de faire partie de mon jury de thèse. La participation de chacun d'entre vous a apporté une dimension essentielle à la soutenance de ma thèse, transformant cette épreuve en une expérience formatrice. Votre contribution va bien au-delà de ce moment clé ; elle s'inscrit dans le prolongement de mon parcours de recherche et continue d'influencer ma réflexion et mes aspirations futures.

Je souhaite exprimer ma profonde gratitude à chacun de mes co-auteurs et collègues, Aurélien Bossard, Ludovic Li et Gabriel Shenouda, dont la collaboration a été essentielle à la réalisation de nos différents articles de recherche. Travailler ensemble sur ces projets a été une expérience enrichissante et stimulante, marquée par un partage de connaissances, une créativité et un engagement sans faille.

Je remercie aussi les deux structures qui m'ont permis de réaliser cette thèse. D'abord le DVRC, pour le cadre académique fourni et Seville More Helory pour le contexte industriel. Ce fut un réel plaisir d'évoluer à l'intersection de ces deux contextes. Au sein du laboratoire DVRC, chaque jour passé a été une occasion d'apprendre, d'échanger et de progresser, entouré de personnes d'exception. Je tiens à remercier chacun de mes collègues pour leur esprit de camaraderie, leur aide précieuse et leurs précieux conseils. À mes collègues de Seville More Helory, merci pour votre accueil chaleureux, votre ouverture d'esprit et votre volonté de partager vos connaissances et votre expertise. La collaboration entre le monde académique et l'industrie est cruciale, et travailler à vos côtés a été une expérience enrichissante qui a ajouté une dimension précieuse à ma recherche. Votre perspective pratique et vos retours constructifs ont grandement contribué à l'application concrète de mes travaux.

Je souhaite également exprimer ma profonde gratitude envers mes proches pour leur soutien indéfectible et leur encouragement constant au cours de ces quatre années jalonnées de défis. Leur présence et leur appui moral ont été des piliers essentiels dans mon parcours.

Il est particulièrement important pour moi de souligner la contribution significative de mon frère, Mustapha Ayoub. Son rôle dans l'achèvement de cette thèse a été crucial ; sans son soutien, ce travail n'aurait pas atteint son terme. Les discussions et les moments fraternels que nous avons partagés ont été des leviers de motivation fondamentaux et ont grandement contribué à ma persévérance et à ma réussite tout au long de ma vie.

Résumé

Face à l'accroissement incessant des données textuelles à traiter, les systèmes de Recherche d'Information (RI) doivent faire face à l'adaptation des mécanismes existants pour une sélection pertinente des ensembles documentaires dédiés à un contexte spécifique. Une difficulté prépondérante réside dans la divergence terminologique entre les termes employés dans les requêtes et ceux présents dans les documents. Cette disparité sémantique, particulièrement prononcée pour les termes de sens proches dans les documents issus de domaines spécialisés de grande taille, pose un défi significatif en RI.

Face à ces enjeux, de nombreuses études se sont limitées à l'enrichissement de requêtes via des modèles supervisés, une approche qui s'avère inadéquate pour une application industrielle et manque de flexibilité. Cette thèse propose une alternative novatrice avec un système de recherche non supervisé basé sur des méthodes d'Apprentissage Profond. La plateforme *LoGE* utilise un modèle de langage masqué pour extrapoler des termes associés, enrichissant ainsi la représentation textuelle des documents. Les modèles utilisés sont entraînés préalablement sur de vastes corpus textuels, intégrant des connaissances générales ou spécifiques à un domaine, optimisant ainsi la représentation des documents.

L'analyse des extensions générées a révélé un déséquilibre entre le signal (termes pertinents ajoutés) et le bruit (termes non pertinents). Pour pallier ce problème, nous avons développé *SummVD*, une approche de résumé automatique extractif, utilisant la décomposition en valeurs singulières pour synthétiser l'information contenue dans les documents et identifier les phrases les plus pertinentes. Cette méthode a été adaptée pour filtrer les termes des extensions en fonction du contexte local de chaque document, afin de maintenir la pertinence de l'information tout en minimisant le bruit.

Mots-clés : Recherche d'Information Ad-hoc, Apprentissage Profond, Plongement de mot, Enrichissement de documents, Filtrage, Optimisation, Non supervisé

Abstract

Faced with the incessant growth of textual data that needs processing, Information Retrieval (IR) systems are confronted with the urgent need to adopt effective mechanisms for efficiently selecting document sets that are best suited to specific queries. A predominant difficulty lies in the terminological divergence between the terms used in queries and those present in relevant documents. This semantic disparity, particularly pronounced for terms with similar meanings in large-scale documents from specialized domains, poses a significant challenge for IR systems.

In addressing these challenges, many studies have been limited to query enrichment via supervised models, an approach that proves inadequate for industrial application and lacks flexibility. This thesis proposes *LoGE* an innovative alternative with an unsupervised search system based on advanced Deep Learning methods. This system uses a masked language model to extrapolate associated terms, thereby enriching the textual representation of documents. The Deep Learning models used, pre-trained on extensive textual corpora, incorporate general or domain-specific knowledge, thus optimizing the document representation.

The analysis of the generated extensions revealed an imbalance between the signal (relevant terms added) and the noise (irrelevant terms). To address this issue, we developed *SummVD*, an innovative extractive automatic summarization approach, using singular value decomposition to synthesize the information contained in documents and identify the most pertinent phrases. This method has been adapted to filter the terms of the extensions based on the local context of each document, thereby maintaining the relevance of the information while minimizing noise.

Keywords : Ad-hoc Information Retrieval, Deep Learning, Word Embedding, Document expansion, Filtering, Optimisation, Unsupervised

Table des matières

Liste des tableaux

Table des figures

Première partie

Introduction & contexte

Chapitre 1

Introduction

Avant l'avènement des ordinateurs et de l'informatique moderne, la recherche d'information était une entreprise largement manuelle. Les bibliothèques étaient les principaux centres de stockage de connaissances où devaient se rendre les gens pour consulter des livres, des journaux, des revues et d'autres documents imprimés. Elles utilisaient des catalogues manuels, souvent sous forme de fichiers de cartes, pour aider les chercheurs à trouver des livres spécifiques. De plus, des indexes imprimés étaient utilisés pour repérer des informations dans des ouvrages plus volumineux. La recherche d'informations spécifiques impliquait souvent de passer de longues heures à parcourir des piles de documents physiques. Cela nécessitait une expertise pour localiser efficacement les informations recherchées.

Dans les années 1950 et 1960, avec l'avènement des premiers ordinateurs, la recherche d'information a commencé à évoluer vers des méthodes plus informatisées. Toutefois, la recherche d'information restait de l'ordre de l'échange oral ou papier. Dans les années 1990 s'est opéré un tournant majeur dans l'histoire de la recherche d'information avec l'avènement des moteurs de recherche en ligne. Cette période a vu l'émergence de plusieurs moteurs de recherche qui ont révolutionné la façon dont les gens trouvaient des informations en ligne.

La recherche d'information en informatique ne débute pas avec le Web. Le but étant de développer des approches permettant de trouver du contenu divers, notamment des publications scientifiques et des fonds documentaires. Ensuite, certaines professions telles que les avocats, les journalistes, ou encore les médecins se sont approprié ces outils de recherche. La recherche d'information a évolué avec ces contextes, et une grande partie de celle-ci continue toujours de traiter de l'accès à l'information non structurée dans divers domaines corporatifs et gouvernementaux.

La recherche d'information a pris une grande ampleur avec le Web. En effet, le volume et la diversité des contenus ont explosé. La recherche a pu se servir des annotations et des analyses sur le comportement des utilisateurs sur les moteurs de recherche pour trouver rapidement une information de pertinence pour l'utilisateur.

Les moteurs de recherche ont plusieurs objectifs auxquels ils doivent faire face :

— Gérer le vocabulaire énorme pouvant atteindre des millions à l'échelle du Web ;

— Traiter la langue introduisant une sémantique dans la recherche ;

— Indexer le contenu pour l'efficacité ;

— Produire un score pour chaque document, afin d'ordonner les résultats par pertinence ;

— Évaluer la qualité du moteur de recherche ;

— Modéliser le Web pour déterminer l'importance d'un document ;

— Intégrer des modèles d'apprentissage automatique de la langue pour vectoriser les documents.

Cette thèse s'intéresse à proposer une nouvelle approche dans ces domaines dans un contexte spécifique.

1.1 Contexte

Cette thèse a été financée par `Seville More Hélory`[1] dans le cadre de la chaire *LegalCluster*[2] au DVRC[3]. La plateforme `LegalCluster`[4], développée chez Seville More Hélory, a différents buts auprès de ses clients :

— Inscrire la digitalisation des fonctions juridiques dans le schéma directeur du système d'information des clients ;

— Gérer et valoriser la donnée juridique au travers d'une logique de plateforme (`LegalCluster`) ;

— Optimiser l'efficacité opérationnelle des fonctions juridiques, éthiques et de conformité ;

— Conduite du changement pour la transformation digitale ;

— Sécuriser les données juridiques et leur conformité.

Dans ce cadre, la thèse s'intéresse à fournir un moteur de recherche pertinent dans un contexte flexible. La plateforme `LegalCluster` propose de définir des "*clusters*" juridiques dans une approche

1. Seville More Hélory : `https://www.business.legalcluster.com/equipe-legalcluster`
2. Chaire LegalCluster : `https://www.devinci.fr/research-center/chaires-dentreprises/chaire-legalcluster-intelligence-juridique`
3. Laboratoire DVRC du Pôle Léonard de Vinci : `https://www.devinci.fr/research-center/`
4. `https://legalcluster.com`

écosystémique, pour garantir la pérennité des informations produites par les acteurs juridiques. Un "cluster" est un regroupement d'acteurs de services juridiques différents ou similaires travaillant sur un projet possédant un contexte juridique lié à un ou plusieurs domaines. Le contexte juridique repose sur plusieurs types de données comme les contrats, les textes de loi, le profil des avocats, le profil des entreprises, etc. Ainsi, `LegalCluster` cherche à offrir un service de recommandation juridique aux cabinets d'avocats, à la direction et aux clients des fonctions juridiques, localement, dans chaque cluster ou à plus grande échelle.

Toutefois, un moteur de recherche dédié à chaque cluster risquerait d'être extrêmement chronophage à construire afin de répondre à chaque contexte local. D'autant que les données à traiter dans un tel contexte sont extrêmement hétérogènes, aussi bien en contenu qu'en taille. En effet, les documents peuvent contenir des contrats, des profils (entreprise ou juriste), des procès-verbaux avec des structures et des tailles très différentes (simple résumé à plusieurs centaines de pages).

De même, les requêtes effectuées par les utilisateurs ont également des tailles variables, pouvant contenir des textes longs (contrats) ou des profils. Cette différence majeure au niveau des requêtes est particulière par rapport aux approches traditionnelles traitant majoritairement de l'inclusion des termes de la requête dans le document. Dans ce cas, la requête peut être aussi grande que le document, voire supérieure. La notion de pertinence des mots de la requête prend alors un poids énorme dans les techniques de Recherche d'Information traditionnelle.

Cette hétérogénéité et cette variation de taille posent de nombreux problèmes aux méthodes classiques et récentes de recherche d'information, aussi bien en termes de pertinence qu'en temps de calcul.

Un problème particulier lié à ce type de domaine est la spécificité du vocabulaire et de la sémantique des documents du corpus. Cette particularité liée aux documents demande un travail supplémentaire de spécialisation du moteur de recherche pour être capable de gérer la pertinence des documents.

Ainsi, l'enjeu principal de cette thèse est de proposer une méthodologie unique capable de s'adapter à un contexte local, avec des représentations de documents comparables, le tout pouvant passer à l'échelle. Par conséquent, cette thèse rapproche deux domaines de recherche et tente de les combiner de manière pertinente et efficace : le traitement automatique du langage naturel (TALN) et la recherche d'information (RI).

Ce document se focalise sur la conception d'un moteur de recherche dédié au contexte spécialisé du corpus, traitant des corpus de grands documents et répondant à des requêtes longues.

Le traitement de texte appliqué aux documents juridiques (ou autres : documents médicaux, publications scientifiques, etc.) reste une tâche complexe souvent liée à des connaissances d'experts du domaine. La complexité de l'objectif peut être résumée [NW17] par ses trois composantes : le langage, la syntaxe et la sémantique.

Ainsi, les documents spécifiques se basent sur un lexique très riche et distinctif qui diffère dans chaque sous-domaine (ici, droit des affaires, droit pénal, etc.), mais qui va aussi s'appuyer sur un langage simple lors de description de faits par exemple.

1.2 Motivations de la thèse

Les approches classiques en traitement automatique du langage, se basant sur des mesures de co-occurrences (TF-IDF, BM25, ontologies, etc.), montrent certaines limites lors de l'élargissement des corpus et documents traités à plusieurs niveaux : la pertinence des similarités, la prise en compte d'informations sémantiques et le manque d'adaptabilité à des domaines spécifiques. En Recherche d'information, la "Recherche Ad-hoc" permet l'association de documents d'un corpus à une requête donnée et s'apparente à la création d'un moteur de recherche. Cependant, dans une application à un domaine spécifique, tel que le domaine juridique, et dans un contexte de plateforme, plusieurs contraintes doivent être prises en compte.

Je présente ci-dessous les trois termes incontournables de la RI :

Terminologie La Terminologie fait le lien entre des termes/notions et leurs définitions dans un domaine spécifique. Dans notre cas d'usage, cela revient à traiter le lexique de domaines spécialisés au sein d'un même corpus. En effet, il est possible d'avoir des termes ne possédant pas le même sens au sein des différents clusters créés, ce qui peut affecter la pertinence en renvoyant des documents trop éloignés du domaine initial.

Cette problématique ouvre la piste à des méthodes statistiques évoluées qui apportent un aspect contextuel aux documents plus précis. Celles-ci reposent sur des auto-encodeurs reposant sur le taux de correspondance des requêtes avec les documents produisant de meilleurs résultats [Pfe+18 ; Ryg+17].

Synonymie La Synonymie définit les liens entre des termes/expressions ayant un sens similaire

ou proche. Dans le cadre de la Recherche Ad-hoc, le décalage de vocabulaire entre la requête et le document reste un problème récurrent. En effet, si l'utilisateur emploie des termes synonymes aux termes du vocabulaire du document, la correspondance entre le document et la requête ne se fera pas.

De plus, l'hétérogénéité des corpus, présent dans des projets spécifiques liant plusieurs domaines, impacte la pertinence des résultats lors de calculs de similarités entre documents, la rendant disparate. En effet, la conséquence peut être une surreprésentativité de l'occurrence interne au document au détriment de l'intérêt des mots dans l'ensemble du corpus. En d'autres termes, les méthodes syntaxiques apportent une pertinence globale plutôt que locale, peu propice à un contexte spécifique.

Performances Lors du développement d'un moteur de recherche destiné à une utilisation très importante et quotidienne, il est important de garder à l'esprit l'importance des performances du côté utilisateur. En effet, l'utilisation d'un moteur de recherche repose sur la durée de réponse de celui-ci pour chacune des requêtes. Si la pertinence de la réponse est ciblée au mépris de la vitesse de réponse, le moteur perdra son intérêt du point de vue de l'utilisateur. Il faut donc limiter le plus possible les traitements exécutés au moment de la recherche et privilégier les processus pouvant être réalisés en amont de celle-ci.

Dans le but de déployer un moteur de recherche pouvant s'adapter à des domaines multiples tout en assurant la pérennité de sa pertinence et de son efficacité, il est nécessaire de s'intéresser aux problèmes sémantiques récurrents. En effet, la terminologie liée à chacun des domaines ainsi que la synonymie des termes du vocabulaire, entre la requête de l'utilisateur et les documents stockés, doivent être prises en compte. De plus, il est essentiel de considérer l'expérience utilisateur en matière de performance du moteur de recherche. Il faut ainsi appliquer le maximum d'opérations avant la réception d'une requête. Notre objectif est donc de développer un algorithme de recherche répondant à ces éléments et augmentant la pertinence des résultats en sortie.

1.3 Problématique

Dans le cadre du développement d'un moteur de recherche ad-hoc pour un corpus de documents textuels longs et spécialisés, diverses complexités doivent être prises en compte.

Le phénomène du décalage de vocabulaire dans la recherche ad-hoc fait référence à la divergence

entre les termes employés dans les requêtes des utilisateurs et ceux utilisés dans les documents au sein d'un corpus spécifique. Cette divergence résulte de la variabilité des expressions linguistiques, de l'utilisation de synonymes, de la diversité terminologique et de la complexité inhérente au langage utilisé dans les domaines spécialisés. Ce décalage de vocabulaire pose un défi majeur aux systèmes de recherche d'information en diminuant la probabilité de corrélation entre les requêtes et les documents pertinents, ce qui affecte négativement la pertinence et la précision des résultats obtenus. En effet, d'une part, des documents pertinents peuvent ne pas contenir les termes requis, faisant en sorte que le signal capté par le moteur de recherche est insuffisant pour capter la pertinence produite pour chaque document. Ainsi, ce décalage de vocabulaire génère un problème de signal lors de la recherche d'information qui constitue une première cible à traiter dans cette thèse.

Une manière de traiter ce problème est l'enrichissement de documents ou de requêtes. Celui-ci inclut l'intégration d'informations contextuelles, l'annotation sémantique, ou l'insertion de résumés, pouvant offrir une dimension supplémentaire à la recherche ad-hoc. En enrichissant les documents, on peut améliorer significativement la pertinence et la précision des résultats de recherche, en fournissant des points d'ancrage supplémentaires pour la correspondance des requêtes. Cependant, cet aspect reste sous-développé dans l'état de l'art actuel. Les progrès récents dans l'amélioration de la recherche ad-hoc se sont principalement focalisés sur l'enrichissement de requêtes. Cette méthode, considérée comme moins onéreuse en raison de la brièveté relative des textes traités, se heurte cependant à des limitations quant à son applicabilité dans le secteur industriel. En effet, les processus de calcul associés à cet enrichissement sont exécutés en temps réel, du côté de l'utilisateur, entraînant ainsi un ralentissement notable de la performance des moteurs de recherche. En revanche, l'enrichissement de documents propose une méthode plus stable et cohérente. Une fois qu'un document a été enrichi hors ligne, les informations ajoutées demeurent pertinentes et utiles pour toutes les requêtes subséquentes. Cette caractéristique contraste avec l'enrichissement de requêtes, qui nécessite une répétition de l'opération pour chaque nouvelle interaction. Toutefois, les deux méthodologies, tant l'enrichissement de documents que celui de requêtes, font face à une problématique commune : la gestion efficace du bruit généré lors du processus d'enrichissement. Cette difficulté réside dans la nécessité d'équilibrer l'ajout d'informations utiles tout en évitant l'introduction d'éléments non pertinents qui pourraient obstruer ou dénaturer le contenu originel.

Ainsi, une problématique globale à laquelle doit faire face un moteur de recherche dans ce cadre de

décalage de vocabulaire est le rapport entre le manque de signal et le bruit généré par l'enrichissement. Cette thèse propose une approche qui traite ce problème grâce à des représentations de texte adaptées au domaine considéré.

Dans notre contexte, un autre enjeu majeur est de traiter les documents de grande taille qui amplifient le problème de signal et de bruit. En effet, l'extraction et la mise en évidence de l'information pertinente est difficile dans un contexte où la redondance et la variété des informations peuvent poser problème. Ainsi, il est crucial de distinguer les données essentielles des éléments moins importants ou répétitifs pour améliorer la pertinence des résultats de recherche. Cet aspect est d'autant plus important dans l'enrichissement de texte. L'identification précise des informations essentielles, des sections nécessitant un enrichissement et des aspects du document qui bénéficieraient d'un développement ou d'une clarification constitue une étape cruciale pour garantir un processus d'enrichissement ciblé et efficace. Cette démarche méthodique permet de prévenir l'accumulation superflue de données et oriente les efforts vers l'augmentation de la pertinence et de la valeur pratique du document en relation avec les requêtes des utilisateurs. Cette approche stratégique assure que l'enrichissement contribue substantiellement à l'optimisation du contenu du document, en accord avec les besoins spécifiques et les intentions des utilisateurs dans le contexte de la recherche d'information ad-hoc.

Dans le panorama actuel de la recherche, la tâche de détection de termes importants est fréquemment associée au domaine du résumé automatique. Cette association découle de la nécessité d'extraire des éléments significatifs au niveau des mots, des phrases, ou même de paragraphes entiers, et de synthétiser ces informations en un document condensé. Cependant, il est à noter que bon nombre des méthodes reconnues dans ce domaine exigent une consommation de ressources considérable en matière de performances. De surcroît, l'impératif d'entraînement préalable des modèles de façon supervisée constitue un obstacle majeur à leur application dans un contexte industriel. Par conséquent, le développement d'une méthode non supervisée, alliant un passage à l'échelle suffisant pour identifier efficacement les informations pertinentes et réaliser des résumés de textes, s'avère impératif. Cette approche innovante permettrait une application plus large et flexible dans diverses situations industrielles, sans les contraintes liées aux méthodes traditionnelles.

Ainsi, cette thèse s'intéresse à traiter le problème de la recherche d'information ad-hoc pour un corpora de documents spécialisés longs. Ce contexte impose un problème de signal-bruit entre les

requêtes et les documents pertinents. Nous proposons donc une approche de résolution reposant à la fois sur l'extension de documents et le résumé de documents.

1.4 Approche

Afin de mieux appréhender les problèmes de traitement du signal textuel, je débute la présentation de mes travaux en détaillant la méthodologie *SummVD* dans le chapitre 2, une technique de résumé automatique et extractif non supervisé qui sera utilisée lors de notre approche de recherche d'information dans le chapitre 3. L'objectif principal de *SummVD* est de synthétiser les informations cruciales d'un document en projetant son vocabulaire dans un espace vectoriel. Cet espace est élaboré à partir d'un modèle de plongement lexical qui, après un apprentissage préalable, génère un ensemble de vecteurs de mots pour le document.

SummVD utilise la décomposition en valeurs singulières (SVD) pour réduire la dimensionnalité des plongements de mots, limitant ainsi la portée du vocabulaire et, par conséquent, minimisant le bruit. Cette stratégie améliore l'efficacité du processus de synthèse textuelle. Au sein de cette approche, les thèmes principaux du document sont identifiés via un partitionnement des vecteurs de mots obtenus après la SVD, représentant le "signal" du document, autrement dit les concepts les plus pertinents. Des scores de pertinence sont ensuite calculés pour chaque phrase du document, en fonction de la présence et de la fréquence des termes appartenant aux thèmes clés identifiés. Une sélection est alors effectuée parmi ces phrases pour constituer le résumé final.

L'emploi de modèles pré-entraînés, combinés à une méthode matricielle de réduction de la dimensionnalité vectorielle, permet de diminuer significativement le temps de calcul, souvent élevé dans les approches existantes. Cette combinaison confère à *SummVD* un caractère non supervisé, simplifiant ainsi son application dans divers contextes. En résumé, *SummVD* se distingue par sa capacité à offrir une méthode de résumé automatique efficace, tout en réduisant la charge calculatoire, ce qui la rend particulièrement adaptée à un large éventail d'applications.

Dans un deuxième temps, ce travail se penche sur les défis récurrents associés à la recherche d'information ad-hoc, notamment le décalage de vocabulaire entre la requête et le document, mais également la gestion de documents longs et un vocabulaire spécialisé. Pour aborder cette problématique avec trois dimensions à traiter, nous introduisons dans le chapitre 3, `LoGE` , une solution d'enrichissement

de documents et de filtrage des extensions qui a pour but de réduire le décalage de vocabulaire spécifique pour de longs documents. `LoGE` se distingue également par sa structure modulaire adaptable à divers contextes industriels.

La solution proposée combine une méthode de recherche classique, BM25, avec un modèle d'apprentissage profond, BERT (*Bidirectional Encoder Representations from Transformers*), reconnu pour son efficacité dans de multiples tâches de Traitement Automatique du Langage Naturel (TALN). Ce modèle non supervisé, qui nécessite uniquement des données textuelles pour son entraînement, est capable de saisir les concepts globaux de domaines spécialisés tels que la médecine ou le juridique. Grâce à de nombreux modèles pré-entraînés sur des corpus étendus, une compréhension approfondie de ces domaines est assurée. Les modèles masqués de BERT permettent de prédire des termes en tenant compte du contexte local des documents. Nous utilisons cette propriété pour générer des extensions de documents après un filtrage initial, visant à concentrer la prédiction sur les termes pertinents.

Cependant, bien que l'enrichissement lexical soit efficace, la présence de bruit (termes non pertinents ajoutés) peut perturber le signal (termes pertinents). Pour résoudre ce problème, nous adaptons la méthode *SummVD*, qui opère au niveau des phrases, en la recentrant sur les termes seuls. Cette adaptation du filtrage des extensions produit des résultats surpassant les méthodes comparables de l'état de l'art. Par ailleurs, l'indexation des termes filtrés améliore l'efficacité et l'explicabilité du modèle pour l'utilisateur, évitant l'effet de « boîte noire ».

Notre approche vise une application étendue à divers contextes. Contrairement aux méthodes actuelles, nous cherchons à minimiser l'impact du moteur de recherche sur l'utilisateur et sur l'entreprise qui l'implémente. Pour ce faire, outre la méthode d'enrichissement, nous proposons une architecture modulaire et distribuée. Cette architecture se caractérise par sa flexibilité remarquable, permettant une permutation aisée des modèles de langage et des filtres. Cette adaptabilité est essentielle pour répondre efficacement aux exigences spécifiques de divers contextes d'application. De plus, elle favorise une répartition stratégique des temps de calcul, en les déplaçant vers la base de données. Cette approche optimise non seulement l'efficacité du traitement des données mais contribue également à une gestion plus rationnelle des ressources calculatoires, alignée sur les besoins et les contraintes de chaque environnement spécifique où le système est déployé.

Pour résumer, les contributions de cette thèse sont :

— La création d'une méthode d'enrichissement lexical non-supervisée de documents longs spécialisés utilisant des modèles d'apprentissage profond pré-entraînés prenant en compte le contexte global du corpus ainsi que le contexte local du document afin de réduire de façon pertinente le décalage de vocabulaire dans une recherche ad-hoc tout en gardant l'intégrité de l'information initiale.

— Le développement d'une méthode de résumé automatique extractif et non-supervisé de document synthétisant les thématiques globales d'un document afin de réduire le bruit et les redondances dans des documents de taille variable en se focalisant sur les informations pertinentes.

— La mise en place d'une architecture modulaire d'un moteur de recherche permettant une intégration simple de nombreux modèles de langage en vue d'enrichir, de filtrer et de rechercher des documents d'un corpus.

Pour finir, le chapitre 4 conclut mon approche en rappelant les objectifs principaux de cette thèse et les solutions apportées à ces problématiques, faisant un bilan des résultats obtenus. Nous aborderons aussi les perspectives de recherche possibles découlant de cette thèse.

1.5 Articles publiés

Durant cette thèse, cinq articles scientifiques ont été publiés dans un journal international (IJWIS), une conférence internationale (IJCNLP), et deux conférences/ateliers nationaux (BDA et EGC) et un poster au GdR TAL :

[ART20] Oussama Ayoub, Christophe Rodrigues et Nicolas Travers. « Adaptive Search Engine for Heterogeneous Documents ». In : *Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA'20)*. 4685. Paris, France, oct. 2020. URL : `https://easychair.org/publications/preprint_open/wFSS`

[ART22] Oussama Ayoub, Christophe Rodrigues et Nicolas Travers. « Un générateur d'extension de documents non supervisé pour moteurs de recherche ». In : *GdR Traitement Automatique de la Langue*. CNRS - IRISA. Rennes, 1er oct. 2022. URL : `https://gdr-tal-rennes.sciencesconf.org/resource/page/id/2`

[She+22] Gabriel SHENOUDA et al. « SummVD : An efficient approach for unsupervised topic-based text summarization ». In : *International Joint Conference on Natural Language Processing (IJCNLP'22), online*. Nov. 2022, p. 501-511. URL : https://aclanthology.org/2022.aacl-main.38 (CORE B)

[Ayo+23] Oussama AYOUB et al. « LoGE : Expansion Locale-Globale de document non supervisée avec un moteur de recherche Extensible ». In : *TextMine Workshop @ EGC'23*. Lyon, France, 2023, p. 41-44. URL : https://textmine.sciencesconf.org/data/pages/TextMine23.pdf

[ART23] Oussama AYOUB, Christophe RODRIGUES et Nicolas TRAVERS. « LoGE : an unsupervised local-global document extension generation in information retrieval for long documents ». In : *International Journal of Web Information Systems* 19.5/6 (nov. 2023), p. 244-262. DOI : 10.1108/IJWIS-07-2023-0109

Deuxième partie

Articles

Chapitre 2

SummVD : une approche non supervisée efficace pour le résumé automatique de textes

Article publié [She+22] :

Gabriel SHENOUDA et al. « SummVD : An efficient approach for unsupervised topic-based text summarization ». In : *International Joint Conference on Natural Language Processing (IJCNLP'22), online*. Nov. 2022, p. 501-511. URL : https://aclanthology.org/2022.aacl-main.38 (CORE B)

Dans ce chapitre, j'introduis *SummVD*, une méthode novatrice pour la synthèse extractive automatique non supervisée. Le problème est de résumer de grands documents, et ce de manière non supervisée pour pallier les problèmes de dimension qui impactent la qualité et les performances de recherche d'information. Ainsi, nous abordons ce problème en résumant le contenu textuel. La plupart des techniques ne passent pas à l'échelle pour de longs documents, car trop gourmandes en ressources, et reposent sur des corpus pré-entraînés ne facilitant pas l'adaptabilité d'un corpus à l'autre - cluster LegalCluster dans notre cas.

Résumé. *SummVD* repose sur la décomposition en valeurs singulières, cette approche linéaire en fonction du nombre de mots vise à diminuer la dimensionnalité des plongements de mots, offrant ainsi une représentation concise des mots sur un nombre restreint de dimensions, chacune incarnant un thème latent. La méthode intègre également le regroupement de mots pour limiter l'étendue du vocabulaire. Cette représentation, spécifique à chaque document, atténue le bruit causé par des dimensions

superflues des plongements dans un contexte limité. Elle est complétée par une heuristique linéaire d'extraction de phrases, conférant à *SummVD* une efficacité notable pour la synthèse de texte.

Par ailleurs, *SummVD* se caractérise par sa faible consommation en ressources, tant en données qu'en complexité, lui permettant d'être appliquée aussi bien à des documents uniques étendus, tels que des articles scientifiques, qu'à d'importants corpus multi documents, et sa rapidité la rend apte à une utilisation dans des systèmes de synthèse en temps réel.

Cette approche sera ensuite évaluée sur divers corpus (actualités, articles scientifiques, réseaux sociaux), surpassant en performance les méthodes extractives récentes. Les résultats nous permettront de mieux appréhender les problèmes de dimension des documents, qui par la suite, nous faciliterons les méthodes de recherche d'information.

2.1 Introduction

Récemment, la recherche en synthèse automatique de textes s'est largement orientée vers des méthodes supervisées. Depuis l'introduction du *Pointer Generator* [SLM17], le champ de la synthèse générative supervisée a connu des développements majeurs [Zho+20 ; Zha+20a ; Wu+21a ; Liu+21]. Toutefois, ces techniques nécessitent des corpus d'apprentissage conséquents, composés d'un vaste ensemble de documents et de leurs résumés associés. Malgré les avancées récentes en matière d'ajustement fin et d'apprentissage par transfert, ces approches demeurent restreintes à des domaines spécifiques. Ainsi, l'exploration des méthodes de synthèse non supervisées demeure cruciale. Dans ce travail, nous nous penchons sur la problématique de la synthèse extractive non supervisée, qui consiste à sélectionner et agencer des phrases issues d'un ou de plusieurs documents afin de créer un résumé. Cette méthode d'extraction repose fréquemment sur les concepts de centralité et de diversité : l'importance centrale d'une phrase dans le texte source et la représentation des informations clés dans le résumé produit.

S'inspirant des travaux de Gong et al. (2018) [Gon+18] sur le calcul de similarité pour les textes longs, nous postulons que des thèmes sous-jacents propres à un texte peuvent se manifester à travers des plongements de mots élaborés à partir d'un corpus général. Chaque thème incarne un aspect distinct de la sémantique du texte. L'identification de ces thèmes sous-jacents facilite l'élimination des éléments superflus dans les représentations lexicales et peut être envisagée comme une représentation

renouvelée du texte. Les mots peuvent être alignés avec un thème sous-jacent, ce qui nous permet de calculer des scores de centralité des mots à partir d'un texte initialement représenté en tant que matrice de plongement de mots. Sur la base de ces scores, une heuristique d'extraction de phrases peut être mise en œuvre pour créer un résumé extractif.

Nous proposons une nouvelle méthode efficace pour la synthèse extractive non supervisée, nommée *SummVD*, dont le code est accessible en ligne [1]. La section 2.2 détaille les méthodes non supervisées récemment développées. Notre méthode est ensuite décrite dans la section 2.3.1. La section 2.4 dévoile nos expériences réalisées sur un éventail étendu de corpus de synthèse, englobant des benchmarks de documents uniques et multi documents, pour évaluer sa capacité de généralisation. Les résultats, présentés dans la section 2.5, surpassent la plupart des méthodes non supervisées récentes sur de nombreux corpus évalués et se rapprochent, dans certains cas, des performances des méthodes supervisées. La section 2.6 aborde la complexité et le passage à l'échelle de notre méthode. La faculté de *SummVD* à traiter efficacement des documents longs et multi documents la rend particulièrement adaptée à la synthèse de divers types de documents, tels que les articles scientifiques.

2.2 État de l'art

2.2.1 Résumé extractif

L'étude de la synthèse extractive remonte à la fin des années 1950, avec les travaux pionniers de Luhn (1958) [Luh58]. Diverses approches, telles que les méthodes symboliques [Edm69], sémantiques [BME99] et statistiques [RJB00], ont été efficacement mises en œuvre pour la synthèse extractive automatique. Des techniques avancées comme l'optimisation linéaire en nombres entiers [GF09] et les algorithmes évolutionnaires [BR17] ont aussi été adaptées à cette fin.

TextRank [MT04], en tant que méthode de synthèse, est fréquemment adoptée comme référence standard. Reposant sur une représentation du texte sous forme de graphe où chaque nœud représente un mot, cette méthode extrait les phrases en se basant sur la centralité de leurs mots au sein du graphe de document.

À notre connaissance, les travaux de Padmakumar et He [PH21] se distinguent comme l'un des processus de résumés extractifs non supervisés les plus novateurs. Leur étude empirique démontre une

1. `https://github.com/SummVD/SummVD`

supériorité sur les approches avant-gardistes dans divers genres textuels, telles que les actualités, le domaine médical et les discussions. Leur modèle s'aligne sur le modèle de similarité de requête décrit par Manning et al. (2008) [MRS08] dans le contexte de la recherche d'information, où un modèle linguistique sert à évaluer la probabilité d'un document à partir d'une requête spécifique. Dans leur méthode, la requête est substituée par une phrase candidate à l'extraction dans le résumé. Ainsi, par une démarche itérative, les phrases sont incorporées au résumé selon les probabilités estimées par le modèle linguistique. Le modèle linguistique employé, GPT-2, est spécifiquement ajusté pour chaque jeu de données afin d'optimiser les résultats. Tous les hyperparamètres sont calibrés sur un échantillon aléatoire de 200 paires document-résumé, visant à perfectionner la mesure ROUGE F1. Cette optimisation concerne le coefficient de pertinence et de redondance dans leur formule d'évaluation des phrases, ainsi que le nombre de phrases à sélectionner pour les méthodes extractives.

SummPip [Zha+20b] représente une méthode de synthèse multi documents non supervisée s'appuyant sur la compression de graphes. Cette technique convertit les documents en graphes de phrases, où chaque nœud symbolise une phrase et les arêtes sont établies sur la base de chaînes lexicales, de marqueurs au niveau discursif, d'informations sémantiques exogènes (WordNet), de références à des entités nommées et d'une similarité sémantique élémentaire basée sur les vecteurs de mots. Cette méthode permet d'intégrer à la fois les représentations linguistiques et les représentations profondes neuronales des documents. Pour générer un résumé de k phrases, un regroupement spectral est utilisé. Pour commencer, une matrice laplacienne est constituée à partir de la représentation graphique des phrases du document, et les k premiers vecteurs propres sont extraits. Chaque phrase se voit ainsi assigner un vecteur de caractéristiques. Par la suite, un regroupement par la méthode des k-moyennes est employé pour segmenter ces phrases en k groupes. La dernière étape implique une compression multiphrase, créant des résumés de documents uniques à partir des groupes découverts. SummPip utilise une version avancée de l'algorithme du plus court chemin pour choisir les phrases définitives du résumé. Pour la partie plongement de mots, un modèle *Word2Vec* [Mik+13], finement ajusté sur chaque jeu de données, est utilisé.

La *Décomposition en Valeurs Singulières* (SVD) appliquée aux textes a d'abord été utilisée pour la comparaison de documents dans le cadre de l'*Analyse Sémantique Latente* (LSA), une technique introduite par Deerwester et al. (1990) [Dee+90]. Dans ce contexte, les documents sont représentés sous forme de matrices document-terme, renseignées par les occurrences de termes dans les documents,

avec un terme par ligne et un document par colonne. La SVD est alors utilisée pour diminuer le volume de termes tout en conservant la similarité entre les documents. Gong et Liu (2001) [GL01] furent les pionniers dans l'application de la LSA à la synthèse automatique de documents. Cette approche permet d'identifier les sujets dominants, puis extrait les phrases les plus proches de ces sujets pour former un résumé. Steinberger et Jezek (2005) [SJ05] ont par la suite affiné cette méthode en ajustant la probabilité de sélection des phrases en fonction de l'importance des sujets, mesurée proportionnellement à leur variance.

2.2.2 Représentation de textes

Plusieurs travaux s'intéressent à la représentation de textes grâce à des techniques de plongement de mots. Je présente ici deux méthodes importantes dans ce domaine.

GloVe [PSM14], signifiant « *vecteurs globaux pour la représentation des mots* », est une technique de plongement de mots qui repose sur un modèle log-bilinéaire avec un objectif basé sur les moindres carrés pondérés. Le concept fondamental de ce modèle est que l'analyse des ratios de probabilités de co-occurrence de mots peut révéler des aspects sémantiques. Il intègre les éléments de deux grandes familles de modèles : la factorisation de matrices globales et les méthodes basées sur les fenêtres de contexte local. Les représentations obtenues révèlent des sous-structures linéaires dans l'espace de vectorisation. Ce modèle a été développé de manière non supervisée à l'Université de Stanford et est disponible en accès libre [2].

BERT pour « *Représentation par Encodage Bidirectionnel à partir de Transformeurs* », créé par Devlin et al. (2019) [Dev+19], représente une avancée récente dans le domaine du pré-entraînement de représentations linguistiques. Il offre des plongements pour les tokens (mots ou sous-mots), mais aussi pour les phrases ou passages de textes. BERT est pré-entraîné à partir de textes non étiquetés, en tenant compte simultanément des contextes gauche et droit dans toutes les couches. Il trouve son application dans une multitude de tâches, incluant la réponse aux questions, l'inférence linguistique, la classification de textes et de phrases, la prédiction de la prochaine phrase, la synthèse de textes, entre autres.

2. `https://nlp.stanford.edu/projects/glove/`

2.3 Notre Approche : SummVD

2.3.1 Modèle proposé

Les plongements de mots offrent une représentation vectorielle des mots basée sur leur contexte. Pourtant, lorsqu'on se concentre sur un contexte précis, comme un document ou plusieurs traitants du même sujet, une grande partie de l'information apportée par ces plongements s'avère superflue, générant du bruit lors des analyses sémantiques. Estimer la similarité sémantique entre deux mots par leurs plongements reste problématique [Far18]. Dans cette optique, nous suggérons d'ajuster des techniques non supervisées pour tirer parti de ces vecteurs et repérer les phrases essentielles d'un texte. Imaginons une matrice où chaque ligne correspond à un mot et chaque colonne à une dimension du plongement vectoriel :

$$Matrice = \#Mot \times \#Dimension.$$

Puisqu'un résumé est en quelque sorte une version condensée d'un texte, notre approche consiste à compresser cette matrice. Notre méthode se décompose en deux phases : réduire d'abord le nombre de mots grâce à une technique de regroupement, puis diminuer les dimensions à l'aide d'une décomposition en valeurs singulières.

La figure 2.1 offre un aperçu de cette approche. Après une étape de pré-traitement, les mots restant dans le document sont vectorisés et projetés dans un espace vectoriel. Des clusters sont alors extraits pour réduire la dimension de l'espace en se concentrant sur les termes représentatifs du document. Ensuite, une réduction de l'espace est appliquée grâce à une SVD, accompagnée d'une fonction de score permettant de sélectionner dans le nouvel espace vectoriel les phrases dont les mots sont les plus représentatifs du document. Je présente dans la suite les détails de l'approche.

2.3.2 Regroupement de mots

Pour minimiser le nombre de mots et leurs vecteurs associés, nous adoptons une technique de regroupement vectoriel non supervisée. Cette approche rassemble les mots qui, une fois vectorisés, semblent partager des contextes similaires dans un même groupe. La technique de regroupement permet d'ajuster le nombre de groupes. Ainsi, un faible nombre de clusters entraîne un taux de compression élevé. Dans un cluster donné, les mots sont remplacés par un vecteur unique qui incarne

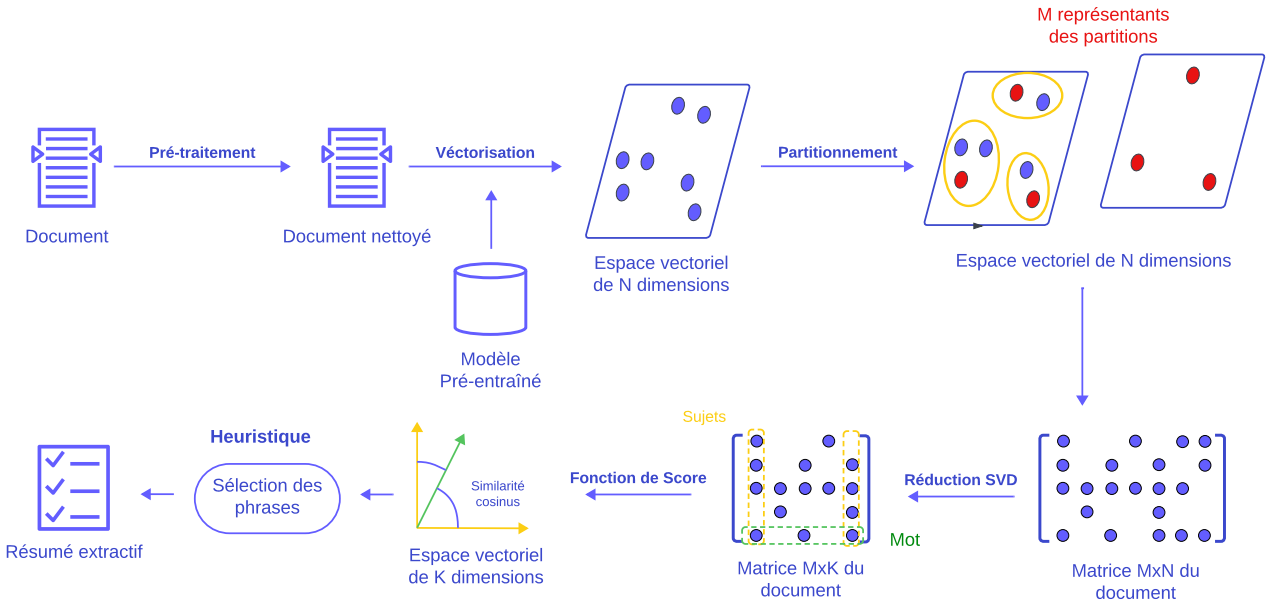

FIGURE 2.1 – Séquence de traitements de notre approche SummVD permettant d'obtenir un résumé extractif à partir d'un document.

ce groupe. Ce vecteur est déterminé en se basant sur sa proximité avec le centre du cluster, parmi tous les vecteurs du groupe.

2.3.3 Décomposition en valeurs singulières (SVD)

Nous suggérons le recours à la SVD pour minimiser les dimensions des plongements de mots. Du fait que ces plongements présentent une grande dimensionnalité (300 dans nos expériences), la SVD est à même de repérer les dimensions essentielles, nous permettant de retenir celles qui sont cruciales. À l'image de la LSA [Dee+90], nous désignons ces vecteurs propres comme des « thèmes ».

Une SVD d'une matrice M de taille $(m \times n)$ est définie comme suit :

$$M = U \cdot \Sigma \cdot V^T$$

avec U une matrice unitaire de taille $(m \times m)$, Σ une matrice diagonale rectangulaire de taille $(m \times n)$ et V une matrice unitaire de taille $(n \times n)$. U et V sont deux matrices orthogonales. U se compose de n vecteurs propres orthonormés liés aux n valeurs propres les plus importantes de MM^T. Quant à V, elle englobe les vecteurs propres orthonormés de M^TM. Σ est une matrice diagonale, avec des valeurs

singulières identifiées comme étant les racines carrées positives des valeurs propres de $M^T M$, classées par ordre décroissant. En ne prenant en compte que les k premières dimensions (où $k < n$), on réduit les dimensions de la matrice M, offrant une approximation utilisable.

2.3.4 Fonction de score des mots

Le score d'un mot pour un thème donné (trouvé par la SVD) est défini par :

$$ScoreMot(\omega, t_i) = \frac{\overrightarrow{\omega} \cdot \overrightarrow{t_i}}{\|\overrightarrow{\omega}\|} \tag{2.1}$$

Où $\overrightarrow{w}$ est le plongement vectoriel du mot w et t_i est un thème identifié par la SVD. Le score est une similarité cosinus entre le plongement du mot et le thème. Intuitivement, plus un mot est proche d'un thème, plus il explique la variation de cet axe, donc plus il contient d'informations et devrait être choisi pour faire partie du résumé.

2.3.5 Extraction des phrases

Nous présentons ici la méthode permettant de sélectionner les phrases les plus pertinentes à partir de la matrice simplifiée, obtenue grâce au regroupement et à la décomposition. Selon l'heuristique de l'algorithme 1, les premiers thèmes dégagés par la SVD servent à identifier une phrase caractéristique pour chaque thème.

Algorithme 1 PhraseParThème(D,k)

Entrée: un document D, nombre de phrases k
Sortie: un résumé res
  $sum = \emptyset$
  pour tout thème k **faire**
    $c = \underset{s}{\operatorname{argmax}}^{s \in D} \frac{1}{|s|} \sum_w^{w \in S} ScoreMot(w, k)$
    $res = res \cup c$
  fin pour

Plus concrètement, pour définir une phrase typique pour chaque thème, l'algorithme 1 choisit la phrase la plus représentative de chaque thème selon la somme des scores de ses mots, cette somme étant ajustée par la longueur de la phrase. La phrase qui correspond le mieux au thème est alors

intégrée au résumé. Cette étape est reproduite pour tous les thèmes. Si le résumé contient k phrases, alors, les k premiers thèmes sont exploités.

2.4 Étude expérimentale

2.4.1 Corpora

Afin d'évaluer notre proposition, nous effectuons l'évaluation sur des corpus hétérogènes. Pour cela, nous comparons notre méthode avec les deux approches non supervisées de résumé extractif les plus proches et récentes à notre connaissance, à la fois sur des tâches de résumé mono et multi documents : PMI [PH21] et SummPip [Zha+20b].

CNN/Daily Mail Introduit par Hermann et al. [Her+15] initialement conçu pour une tâche de réponse automatique à des questions et utilisé pour la première fois pour le résumé automatique par Nallapati et al.[Nal+16]. Ce corpus est composé d'articles de journaux extraits de CNN et Daily Mail. Chaque article est associé à un résumé construit en concaténant les points forts de l'article défini par son auteur. Sa grande échelle permet de l'utiliser dans des méthodes de résumés génératifs neuronaux. La version que nous utilisons est la version non anonymisée.

XSum Le jeu de données XSum pour Extreme Summarization a été introduit par Narayan et al. [NCL18] pour évaluer les systèmes de résumé de documents uniques. Les documents sont collectés à partir des articles de la BBC (de 2010 à 2017). Chaque article est associé à un résumé d'une seule phrase, plus précisément la phrase d'introduction qui le précède, rédigée professionnellement par l'auteur de l'article.

PubMed [Coh+18] Il s'agit d'un jeu de données à document unique composé principalement d'articles scientifiques médicaux associés à leur résumé. Il s'agit de documents longs avec un vocabulaire spécifique.

Reddit [OCM17] Est un ensemble de données reposant sur Reddit composé de 476 récits personnels qui sont utilisés comme documents sources pour le résumé. Ces récits proviennent de 19 forums différents et sont associés à deux résumés de référence : un résumé abstractif et un résumé extractif, tous deux rédigés à la main par quatre étudiants. Nous utilisons le même ensemble de tests que dans cet article [PH21], soit 48 exemples choisis au hasard.

Multi-News [Fab+19] Est un ensemble de données de résumés d'articles de journaux multi documents. Les articles sont extraits de ce site [3]. Comme la majorité des méthodes de résumé de texte utilisent la forme tronquée du corpus, nous faisons de même afin de pouvoir nous comparer.

DUC 2004 Créé pour la campagne d'évaluation de la synthèse documentaire de la Document Understanding Conference, DUC2004 [Ped+11], cet ensemble de données multi documents comprend 50 ensembles de 10 articles de presse, chaque ensemble abordant un thème spécifique. Chaque ensemble de ces 50 groupes est apparié à un résumé rédigé par un humain. Les articles de chaque groupe sont fusionnés en un document unique, formant ainsi un corpus de 50 documents d'une longueur considérable, chacun étant associé à un résumé standardisé de haute qualité.

2.4.2 Modèles de référence

TextRank Nous avons implémenté TextRank, une approche extrêmement répandue et couramment utilisée dans le domaine de la synthèse de texte. Comme décrit dans la section 2.2, cette méthode représente, à l'heure actuelle, l'une des stratégies non supervisées les plus efficaces en termes de rapidité pour la production de résumés. Pour cela, nous appliquons l'implémentation fournie par Gensim [4] [Bar+15].

LSA Nous appliquons LSA [SJ05], une technique fondée sur la SVD, tel qu'expliqué dans la section 2.2. Cette méthode sert à souligner les bénéfices de notre approche qui repose sur l'utilisation de plongements de mots.

BERT SVD Nous avons conçu une approche originale fondée sur les vecteurs extraits à partir de BERT. Cette méthode offre la possibilité de représenter des phrases ou des passages entiers plutôt qu'uniquement de simples mots. Après la vectorisation de toutes les phrases d'un document, le processus adopté est analogue à celui de notre méthode principale, *SummVD*. En outre, la phase finale de sélection des phrases est effectuée de manière méthodique, en privilégiant les phrases les plus étroitement liées aux thèmes principaux du document.

PMI Nous avons procédé à l'application de PMI [PH21] en utilisant l'implémentation mise à disposition par ces auteurs [5]. Notre utilisation se concentre exclusivement sur les ensembles de données dédiés à la synthèse de documents uniques, étant donné que PMI est spécifiquement conçu pour cette

3. `http://www.newser.com`
4. `https://radimrehurek.com/gensim/`
5. `https://github.com/vishakhpk/mi-unsup-summ`

tâche.

SummPip Nous avons mis en œuvre SummPip [Zha+20b] selon l'implémentation proposée par ses auteurs [6]. SummPip étant élaboré pour la synthèse de documents multiples, notre utilisation s'est strictement focalisée sur les jeux de données impliquant la synthèse de multiples documents.

Supervised représente le modèle MatchSum [Zho+20]. Il constitue l'une des stratégies d'extraction les plus récentes et les plus avancées dans le domaine de l'apprentissage profond supervisé.

2.4.3 Détails d'implémentation

Pour le pré-traitement des données, nous avons employé les outils NLTK [7], éliminant ainsi les mots vides et les caractères spéciaux. De plus, le parseur de phrases NLTK a été utilisé pour identifier les différentes phrases au sein des documents. Dans le but d'effectuer une comparaison directe et équitable entre toutes les approches de synthèse de texte non supervisée et notre méthode, nous avons produit des résumés de longueur équivalente à ceux générés par PMI [PH21] et SummPip [Zha+20b], mesurés en nombre de phrases. Pour les ensembles de données CNN/DM et XSum, nous avons opté pour 3 phrases ; pour Reddit, 4 phrases ; pour PubMed et MultiNews, 9 phrases ; et pour DUC 2004, 7 phrases.

Dans le but de préserver la simplicité et l'autonomie de notre méthode en demeurant le plus possible non supervisée, nous avons opté, après évaluation, pour l'utilisation d'une méthode de vectorisation de mots générique pré-entraînée : GloVe (entraîné sur le corpus Common Crawl, constituée de 840 milliards de tokens, d'un vocabulaire de 2,2 millions, sensible à la casse, générant des vecteurs de 300 dimensions). C'est cette méthode qui s'est avérée être la plus efficace.

Nous avons évalué quatre techniques de regroupement : OPTICS [Ank+99], une version améliorée de DBSCAN [Est+96], l'algorithme des K-Means [For65], et le regroupement agglomératif, en utilisant leur mise en œuvre dans la bibliothèque *scikit-learn* [Ped+11]. L'adoption du regroupement agglomératif se traduit par une diminution marginale du score ROUGE, estimée entre 0,5% et 1,3% par rapport aux K-moyennes, et entre 1,0% et 1,9% en comparaison avec OPTICS. Cependant, cette approche offre des avantages notables en termes de rapidité d'exécution, avec des gains variant de 40% à 700% et de 1 100% à 2 300%, en fonction du corpus. Ainsi, le regroupement agglomératif se

6. `https://github.com/mingzi151/SummPip`
7. `https://www.nltk.org/`

révèle être un équilibre judicieux entre performance et rapidité d'exécution, un critère essentiel pour le passage à l'échelle de notre proposition.

Concernant la détermination du nombre de groupes, nous avons recours à la méthode du coude. Cette technique nous permet de déterminer, de manière moyenne et automatique, le nombre de groupes le plus approprié pour chaque corpus.

Nom du corpus	**Nature du document**	**Type**	**Taille d' échantillon**	**Phrases par document**	**Mots par document**	**Phrases par résumé**	**Mots par résumé**	**Taux de compression**
CNN/DM	Presse	mono	11 489	26,9	766,6	3,9	58,2	7,6%
XSum	Presse	mono	11 331	23,2	424,9	1	18,6	4,4%
PubMed	Scientifique	mono	6 658	101,6	3 142,9	7,6	208,0	6,6%
Reddit	Forum	mono	48	12,1	234,5	1,2	25,2	10,7%
Multi-News	Presse	multi	5 622	17,5	491,0	9,8	262,0	53,4%
DUC2004	Presse	multi	50	264,9	6 583,1	31,1	422,3	6,4%

TABLE 2.1 – Caractéristiques du Corpora : taille des échantillons testés (en nombre de documents), nombre moyen de phrases par document, nombre moyen de mots par document, nombre moyen de phrases par résumé de référence et taux de compression (c.f. équation 2.2) pour chaque corpus décrit à la section 2.4.1

.

Le tableau 2.1 présente les caractéristiques de l'ensemble des corpus mentionnés dans cette section et utilisés dans notre procédure d'évaluation. Ce tableau souligne les différences entre ces corpus, notamment en termes de types de résumé (document unique vs multi documents), de la nature des documents (scientifiques, articles de presse, contenus de médias sociaux), des longueurs des documents et des résumés de référence, ainsi que du taux de compression, calculé selon l'équation suivante :

$$CompRate(D, A) = \frac{|A|}{|D|} \tag{2.2}$$

Où D désigne le nombre de mots du document source et A le nombre de mots du résumé correspondant.

2.5 Résultats

La figure 2.2 illustre un exemple de résumés obtenus par les différentes méthodes testées sur un article du corpus CNN/DM. Afin d'évaluer notre méthode, nous avons recours à la mesure ROUGE F1, largement reconnue dans le domaine [Lin04]. Cette méthode consiste principalement à compter le nombre de mots ou groupes de mots contigus en communs entre le résumé candidat généré automatiquement et le résumé de référence généré manuellement. La bibliothèque Python que nous employons

TextRank : a father-of-three and popular radio host in berkeley, california, was killed in a hit-and-run in the early hours of saturday morning. Wesley burton, a father-of-three and popular radio host at kpfa in berkeley, california, was killed in a hit-and-run in the early hours of saturday morning as he drove home from work. burton had three children - santiago, enrique, and samaya – aged between 4 and 9 and after growing up without a father his dream had been to raise his own kids

LSA : the crash occurred near the berkeley-oakland city line and police say the hit-and-run driver fled on foot. a gofundme account has been set up to help burton 's wife pay funeral costs and other family expenses. police are urging anyone with information to call the traffic investigation unit on (510)777-8570.

PMI : a father-of-three and popular radio host in berkeley, california, was killed in a hit-and-run in the early hours of saturday morning. his wife lucrecia has made a tearful plea for anyone with information to come forward and speak to the police. we lost our rock. he was our stability, our strength, ' she told ktvu.

BERT SVD : ‘ help us regain our peace. burton had three children - santiago, enrique, and samaya – aged between 4 and 9. oakland crime stoppers is offering a $ 10,000 reward for information leading to an arrest.

SummVD : a father-of-three and popular radio host in berkeley, california, was killed in a hit-and-run in the early hours of saturday morning. wesley burton, who worked at kpfa, was driving home from work when a white dodge charger crashed into his silver mercury. the crash occurred near the berkeley-oakland city line and police say the hit-and-run driver fled on foot.

FIGURE 2.2 – Exemples de résumés générés pae SummVD et par différentes méthodes de l'état de l'art décrités section 2.4.2 sur à partir d'un même article appartenant au corpus CNN/DM.

est disponible ici [8]. Cette procédure est comparable à l'exécution du script perl ROUGE de la manière suivante :

`"ROUGE-1.5.5.pl -m -e ./data -n 2 -a /tmp/rouge/settings.xml"`.

2.5.1 Résultats ROUGE

Le tableau 2.2 présente les résultats obtenus à l'aide de la métrique ROUGE F1. Nous pouvons observer que *SummVD* obtient des résultats très encourageants comparativement à PMI, SummPip, et TextRank dans une majorité de situations. Bien que notre approche ne s'avère pas constamment supérieure, elle démontre une efficacité comparable tant pour la synthèse de documents isolés que pour des ensembles de documents. De plus, sa performance ne semble pas être influencée par la longueur des documents, une caractéristique cruciale pour la synthèse d'articles scientifiques ou de projets impliquant la concaténation de multiples sources.

À la lumière des deux ensembles de données multi documents analysés, notre stratégie affiche d'excellents résultats par rapport aux autres techniques non supervisées. Une observation minutieuse

8. `https://pypi.org/project/rouge-score/`

	Mono-document												**Multi-document**					
	CNN/DM			**XSum**			**Reddit**			**PubMed**			**Multi-News**			**DUC2004**		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Supervisé	44,41	20,86	-	24,86	04,66	-	-	-	-	41,21	14,91	-	46,20	16,51	-	-	-	-
Lead-k	**40,13**	17,63	**25,09**	19,52	02,67	12,45	25,66	07,51	17,94	37,98	13,55	20,16	42,35	14,14	**20,02**	30,66	08,36	14,73
TextRank	32,87	13,90	20,93	18,67	**03,15**	12,23	26,55	08,64	19,01	36,93	13,60	**20,96**	34,50	10,86	17,42	24,41	08,32	13,44
LSA	29,23	10,47	18,35	18,70	02,60	11,82	25,12	07,74	17,26	33,55	09,00	16,02	32,65	09,22	16,36	22,68	08,09	11,73
PMI	36,56	15,49	23,11	19,13	02,89	12,45	**28,22**	08,51	**20,63**	37,82	10,85	18,33	-	-	-	-	-	-
SummPip	-	-	-	-	-	-	-	-	-	-	-	-	42,32	13,28	-	36,3	08,47	-
BERT SVD	25,28	7,60	15,90	17,09	02,44	11,41	22,14	05,60	14,77	33,85	09,43	16,45	40,86	13,42	18,44	18,57	03,76	10,27
SummVD	39,36	**17,70**	24,70	**19,7**	02,77	**12,70**	28,12	**09,27**	19,07	**38,06**	**14,49**	20,20	**43,55**	**15,83**	19,23	**37,80**	**10,15**	**16,43**

TABLE 2.2 – Résultats ROUGE-1, ROUGE-2 and ROUGE-L F1 pour chaque méthode décrite à la section 2.4.2 et pour notre proposition SummVD décrite à la section 2.3.1. Les meilleures méthodes non supervisées sont représentées en gras.

du tableau 2.2 révèle que la technique supervisée *MatchSum* domine nettement les alternatives non supervisées pour des ensembles caractérisés par des documents de moindre envergure. Toutefois, dans le contexte d'ensembles dotés de documents plus conséquents, tels que PubMed et Multi-News, la distinction entre *MatchSum* et les techniques non supervisées s'amenuise.

Il convient de souligner que, en se basant sur la métrique ROUGE-2, *SummVD* est distingué comme étant le système non supervisé prédominant pour cinq des six ensembles de données. Une étude conduite par Graham et al. [Gra15] a établi que ROUGE-2 était la métrique de l'ensemble ROUGE la plus alignée avec les évaluations subjectives humaines, contrairement à ROUGE-1 et ROUGE-L qui, à l'instar de ROUGE-W, s'avèrent moins adéquates.

2.5.2 Temps d'exécution

Dans le tableau 2.3, une évaluation comparative du temps d'exécution entre *SummVD*, TextRank, PMI et SummPip est présentée. Pour les quatre approches, nous mesurons le temps d'exécution moyen nécessaire pour synthétiser un document.

Lors de cette évaluation, il convient de noter que nous n'avons pas inclus le processus d'adaptation par apprentissage (ou fine-tuning) du modèle de langage associé à PMI et SummPip nécessaire pour chaque corpus à traiter qui est un processus très chronophage. Conformément aux instructions d'utilisation stipulées par les auteurs sur la page GitHub dédiée de ces méthodes, les codes ont été exécutés successivement dans un environnement de travail vierge. La machine utilisée pour toutes ces expériences repose sur un processeur AMD 3700x avec 8 coeurs, 64 Go de mémoire vive, 2 cartes graphiques RTX 2080Ti de 11 Go chacune et utilise le système d'exploitation Windows 10.

	Mono-document				Multi-document	
	CNN/DM	**Xsum**	**Reddit**	**PubMed**	**Multi-News**	**DUC2004**
TextRank	0,02s	0,01s	0,01s	0,09s	0,046s	0,32s
PMI	72,72s	56,28s	25s	448,2s	-	-
SummPip	-	-	-	-	6s	846s
SummVD	0,1s	0,07s	0,05s	0,3s	0,23s	0,52s

TABLE 2.3 – Temps moyen de génération d'un résumé pour chaque méthode décrite à la section 2.4.2 et sur chaque corpus décrit au tableau 2.1

Il est à noter que TextRank s'est avéré être le plus performant des quatre, présentant un temps d'exécution en moyenne cinq fois inférieure à celui de *SummVD*. La rapidité de TextRank est bien documentée, et l'implémentation Gensim que nous employons a été optimisée pour maximiser cette vélocité. Une analyse des tableaux 2.1 et 2.3 révèle que le temps d'exécution de SummPip s'accroît d'un facteur de 141 lorsque le nombre de mots par document augmente de 6,74 (comparaison entre Multi-News et DUC2004). En revanche, l'accroissement du temps d'exécution pour *SummVD* est seulement de 2,2. Ainsi, pour le jeu de données DUC2004, *SummVD* se révèle être en moyenne 1 626 fois plus efficient que SummPip. Une comparaison avec PMI démontre que *SummVD* est en moyenne 885 fois plus efficient sur l'ensemble des jeux de données concernés.

En considérant la densité lexicale, PubMed contient en moyenne 6,74 fois plus de termes que CNN/DM, XSum et Reddit. En conséquence, *SummVD* montre un temps d'exécution 4,28 fois supérieur sur PubMed par rapport aux autres jeux. Pour PMI, ce ratio est de 8,62. Notons que PMI est 1 494 fois moins efficient que notre méthode sur PubMed. À titre de comparaison, la méthode supervisée de référence MatchSum [Zho+20] nécessite un temps d'entraînement de 30 heures pour le corpus CNN/DM, exécutée sur une infrastructure matérielle dédiée munie de 8 cartes graphiques de type V100.

2.6 Discussion

2.6.1 Complexité

À notre connaissance, hormis MMR [CG98] et ses méthodes associées aux performances relatives, il n'existe pas de méthode entièrement linéaire pour générer des résumés extractifs. La complexité algorithmique de la SVD [GVL96] est caractérisée par :

$$O(mn\,min\{n, m\})$$

Dans notre cas, m dénote le nombre de mots et n le nombre de dimensions du plongement de mots. De plus, dans ce contexte précis, bien que la dimension des colonnes soit déterminée par la taille du plongement (soit 300 dans nos expériences), celle-ci demeure constante, indépendamment de la longueur du document. Par conséquent, augmenter la longueur du document introduira uniquement de nouvelles entrées (mots). Ainsi, pour des documents dont le nombre de mots dépasse la taille du plongement de mots, la complexité de la SVD s'établit comme étant quadratique en n et linéaire en m. Puisque n est un paramètre fixe, la complexité de la SVD devient linéaire par rapport au nombre de mots lorsque $m > 300$. Ceci explique pourquoi notre approche passe à l'échelle plus facilement lorsque le nombre de mots à traiter augmente comme illustré à la figure 2.3.

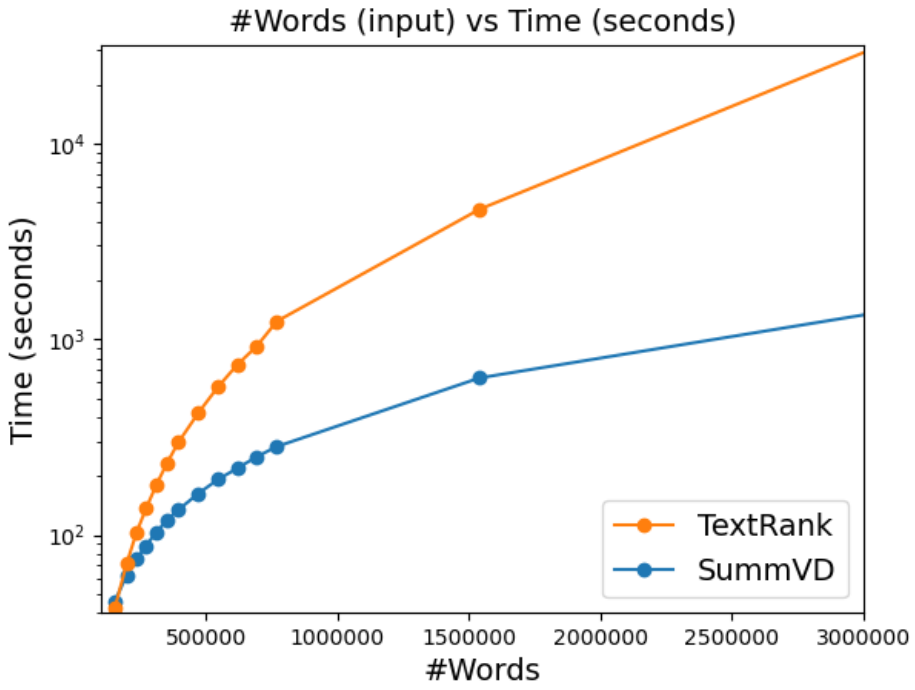

FIGURE 2.3 – Temps moyen pour générer un résumé en fonction du nombre de mots en entrée pour SummVD et TextRank (implémentation de Gensim). Le temps est représenté à l'aide d'une échelle logarithmique.

2.6.2 Passage à l'échelle

La complexité algorithmique de *SummVD*, dépeinte à la figure 2.3 sur une échelle logarithmique, montre la capacité de l'approche à passer à l'échelle de façon significative. Par rapport à l'implémentation de TextRank [Bar+15] présente dans Gensim [ŘS10], on observe un écart notable en matière de temps de calcul concernant les documents de grandes dimensions, *SummVD* se montrant dans ce cas beaucoup plus performant. De ce fait, *SummVD* se positionne comme un outil pertinent pour des opérations telles que la synthèse en temps réel de textes d'envergure, résumé d'informations journalières ou la synthèse d'un ensemble de documents.

2.6.3 Analyse SVD

La SVD constitue l'élément central de *SummVD*. Il est donc primordial de caractériser son impact sur le processus de résumé. Dans l'analyse présentée dans le tableau 2.4, nous quantifions les catégories grammaticales (*Part-Of-Speech tag*) de l'ensemble des termes présents dans les documents sources de chaque corpus évalué, ainsi que les catégories grammaticales de chaque mot dominant suite à l'application de la SVD. Une observation des disparités dans la distribution de ces balises entre ces deux ensembles de termes peut offrir une première indication quant au type de termes que la SVD tend à privilégier.

	NOM	**VERB**	**PROPN**	**NUM**	**ADJ**	X	**INTJ**	**PRON**	**ADP**	**SYM**	**PONC**	**DET**	**ADV**
Source	21,3	12,6	5,5	1,7	6,7	0,2	0,2	6,7	10,9	0,1	7,1	8,1	4
Après SVD	38	25,6	14	2,7	8,3	0,7	0,5	2,4	1,5	0	0,6	0	2,7

TABLE 2.4 – Proportion des principales catégories grammaticales (POS tag) dans les documents sources avant et après réduction par la SVD.

Le tableau 2.4 montre que la distribution des catégories grammaticales dans les documents sources varie considérablement de celle des mots choisis après la SVD. Il s'avère que la SVD privilégie automatiquement les termes les plus informatifs tels que les noms communs, les verbes, les noms propres, et les nombres, tout en omettant ceux jugés moins pertinents comme les prépositions, les adverbes et les interjections, indépendamment de leur fréquence d'apparition. Les catégories grammaticales mises en évidence par la SVD sont représentées en bleu, tandis que celles qui sont minimisées sont représentées en rouge.

2.6.4 Analyse des résultats de BERT SVD

L'adoption de BERT en tant que méthode d'encodage de phrases ne semble pas générer les performances optimales anticipées. Effectivement, en mobilisant la meilleure configuration des couches cachées de BERT dédiée à la synthèse de texte, nous aboutissons aux résultats illustrés dans le tableau 2.2. Cette divergence en comparaison avec le modèle s'appuyant sur GloVe est difficile à expliquer. Néanmoins nous pouvons émettre l'hypothèse que l'efficacité de ce dernier pourrait être attribuée à la capacité de la SVD à discerner le poids sémantique d'un terme précis au sein d'un document.

Par contraste, un terme d'importance pourrait être occulté dans la représentation globale de l'encodage de phrase offerte par BERT. Ces observations suggèrent une perspective notable : les synthèses pourraient être principalement articulées autour de la prééminence de termes distinctifs, que notre approche basée sur la SVD est en mesure de déceler.

2.7 Conclusion

Dans ce chapitre, j'introduis *SummVD*, une méthode fondée sur le plongement lexical et des techniques non supervisées, offrant des résumés à la fois rapides et fiables. J'y ai décrit une heuristique d'extraction capable d'exploiter efficacement la matrice documentaire réduite, applicable tant aux documents uniques qu'aux ensembles multi documents. Cette technique repose essentiellement sur l'extraction des points représentatifs du document grâce aux clusters de représentations vectorielles, d'une réduction SVD de l'espace et d'une sélection de phrases représentatives grâce aux mots ainsi représentés et réduits.

Une évaluation exhaustive a été réalisée sur des corpus diversifiés. L'étude empirique révèle des résultats prometteurs en comparaison avec l'état de l'art, tant en matière d'efficacité selon les scores ROUGE qu'en rapidité de traitement. En comparaison avec les méthodes récentes, *SummVD* affiche en moyenne des scores ROUGE supérieurs, tout en étant environ mille fois plus rapide sur les jeux de données comportant les documents les plus longs.

Cette performance est obtenue sans adaptation spécifique des *plongements de mots* aux différents domaines ; par conséquent, des améliorations sont envisageables dans des domaines utilisant un vocabulaire spécifique, comme le secteur médical, scientifique ou encore les médias sociaux. La flexibilité de *SummVD* vis-à-vis de la nature et de la taille des documents suggère de vastes possibilités d'explo-

ration sur de grands ensembles de données multi documents, à l'image de ceux de *Google, TripAdvisor* ou encore *Amazon*.

Ainsi, grâce à cette approche, nous pouvons réduire la dimension des documents en extrayant les informations représentatives de ceux-ci, et ce dans un temps raisonnable. Nous pouvons maintenant nous intéresser à la manière de retrouver des documents dans un corpus de documents de grandes dimensions. Nous allons donc étudier les techniques de Recherche d'Information et les problèmes de différence de vocabulaire entre les requêtes et les documents. Nous devrons donc à la fois améliorer les documents et les réduire (pour diminuer l'espace des mots extraits des documents) pour à la fois améliorer la qualité des résultats et leur performance.

Chapitre 3

LoGE : une approche non supervisée par extension Locale et Globale de documents pour la recherche d'information dans des documents longs

Articles publiés :

Oussama Ayoub et al. « LoGE : Expansion Locale-Globale de document non supervisée avec un moteur de recherche Extensible ». In : *TextMine Workshop @ EGC'23*. Lyon, France, 2023, p. 41-44. URL : https://textmine.sciencesconf.org/data/pages/TextMine23.pdf

Oussama Ayoub, Christophe Rodrigues et Nicolas Travers. « LoGE : an unsupervised local-global document extension generation in information retrieval for long documents ». In : *International Journal of Web Information Systems* 19.5/6 (nov. 2023), p. 244-262. DOI : 10.1108/IJWIS-07-2023-0109

Afin de concevoir un moteur de recherche intégrant un corpus de documents sur un domaine spécifique (exemple : juridique, médical) avec des documents de grande taille, il est nécessaire de résumer le contenu textuel de ceux-ci. À l'instar de *SummVD* présenté précédemment, le résumé de documents est une première approche permettant d'extraire l'essence même du texte, toutefois, la synthèse produite n'est pas toujours adaptée aux requêtes soumises au moteur de recherche. Ainsi, dans le cadre de la Recherche d'Information, la différence de vocabulaire entre la requête et le document réduit considérablement la qualité des résultats. Je propose donc une approche de Recherche d'Information permettant de pallier ce problème de vocabulaire tout en réduisant la taille des documents produits. Le vocabulaire étendu repose sur des techniques de plongement de mots grâce à un modèle de type

BERT. Cette approche a été développée dans la plateforme `LoGE` qui intègre le corpus "étendu" dans un cluster *Elasticsearch*.

Résumé. Face à l'accroissement incessant des données textuelles à traiter, les systèmes de Recherche d'Information (RI) contemporains requièrent des solutions performantes afin de sélectionner de manière efficiente l'ensemble documentaire le plus adéquat pour une requête spécifique. La divergence terminologique entre les mots formulant une requête et ceux présents dans les documents pertinents constitue une problématique notoire pour les systèmes de RI. Cette disparité sémantique, avec des termes ayant des sens proches, s'avère particulièrement problématique pour les systèmes de RI, surtout lorsqu'il s'agit de documents de grande envergure issus de domaines spécialisés.

Le présent chapitre propose un système de RI non supervisé innovant, s'appuyant sur des méthodes avancées d'Apprentissage Profond (AP). Ce système emploie un modèle de langage masqué par AP pour extrapoler des termes associés, enrichissant ainsi la représentation textuelle du document. Les modèles d'AP mobilisés, préalablement entraînés sur de vastes corpus textuels, intègrent des connaissances générales ou spécifiques à un domaine, optimisant par là même la représentation documentaire.

Notre méthodologie est soumise à une évaluation rigoureuse dans le cadre de domaines de RI spécifiques, traitant de documents de grande taille, afin de montrer la généralisation possible du modèle proposé et les résultats prometteurs qu'il permet d'atteindre.

3.1 Introduction

La Recherche d'Information (RI) a connu une évolution constante sur quatre décennies, passant de représentations symboliques à des représentations vectorielles, et évoluant à travers des transformations et des analyses de texte, ainsi que des traitements hors ligne et en ligne. Dans ce contexte, la recherche ad hoc vise à mettre en avant des documents contenus dans un corpus en relation avec une requête donnée, résumant ainsi les fonctionnalités escomptées des moteurs de recherche actuels. Un défi persistant dans la recherche ad hoc réside dans la discordance de vocabulaire entre les requêtes et les documents, discordance qui peut être caractérisée de deux manières.

La première est une question terminologique, qui associe des termes et notions à leurs définitions dans un domaine donné, ce qui est particulièrement saillant dans les domaines spécialisés tels que la médecine ou le droit, et peut induire l'utilisateur à se focaliser sur des documents non pertinents. La seconde est une question synonymique, qui concerne la relation entre des termes ou expressions de sens similaire. Cette dernière peut omettre des documents qui n'incluent pas les termes utilisés dans la requête, nécessitant ainsi d'effectuer des recherches plus approfondies.

Les méthodes classiques basées sur la cooccurrence de termes, comme le BM25 [RJ76], ne parviennent pas à saisir la sémantique ni à faire correspondre les concepts. Afin de pallier cet écart sémantique entre requêtes et documents, la recherche pour des améliorations sémantiques s'est intensifiée. D'une part, les représentations vectorielles de textes, complétées par des informations sémantiques via des réseaux de neurones, visent à combler les lacunes au niveau des mots, des phrases ou des documents. Ces méthodes, qui transforment les textes en vecteurs, entraînent cependant une perte d'explicabilité et peuvent augmenter le temps de traitement des requêtes en raison de la lourdeur relative de la vectorisation et de la comparaison des vecteurs par rapport aux systèmes de RI standards.

De plus, la supervision prédominante dans la plupart des modèles restreint leur application universelle, les rendant inadaptés à d'autres ensembles de données spécifiques à un domaine. D'autre part, l'expansion des requêtes, qui consiste à ajouter de nouveaux termes pour pallier le déficit de vocabulaire, a été largement explorée. Néanmoins, beaucoup de ces approches se concentrent sur l'enrichissement des requêtes au détriment de la performance temporelle. Il est à noter que peu d'études se sont penchées sur l'expansion des documents. Comme cette expansion peut être effectuée lors de la phase hors ligne, elle rend la phase de requête plus efficiente et précise, en complétant notamment le manque de vocabulaire pour des domaines spécifiques et pour des documents de grande longueur.

Enfin, elle favorise la proposition d'un modèle généralisé et intelligible pour les utilisateurs. Le principal enjeu de cette expansion textuelle est d'équilibrer la pertinence (par l'ajout de synonymes) par rapport au bruit (par l'ajout de termes superflus). Ainsi, une compréhension approfondie de l'impact de l'expansion textuelle et de sa portée est indispensable pour améliorer adéquatement le moteur de recherche.

Dans ce chapitre, nous introduisons un cadre générique, non supervisé et modulaire, LoGE, pour la

RI ad hoc, qui propose une expansion documentaire sur-mesure pour des domaines spécifiques, visant à rehausser la pertinence et à étendre les capacités des moteurs de recherche traditionnels. Ce processus s'appuie sur une génération automatique d'extensions à partir de modèles prédéfinis, débutant par un filtrage du texte pour isoler les mots pertinents, suivi de la proposition de synonymes via des modèles de langage pré-entraînés, et se concluant par un filtrage final basé sur une mesure d'importance des mots au sein des documents.

Notre contribution se synthétise de la manière suivante :

1. Un processus non supervisé, destiné à atténuer la disparité lexicale entre les requêtes et les documents par le biais d'une expansion documentaire,

2. Un cadre modulaire élaboré sur la base des moteurs de recherche traditionnels, incorporant un modèle textuel pré-entraîné externe, tel que BERT, destiné à pallier le déficit sémantique, ainsi qu'un processus de génération de résumés extractifs, inspiré de la technique de Décomposition en Valeurs Singulières (SVD), dans le but de minimiser le bruit engendré dans les documents étendus.

   Ce cadre peut être personnalisé en intégrant conjointement le modèle BERT, adapté à un domaine spécifique, et le résumé, en le combinant avec d'autres modèles basés sur la SVD,

3. Une analyse d'impact exhaustive sur les expansions de documents et une étude ablative de notre processus menée sur deux benchmarks spécialisés : NFCorpus et la législation de l'UE/Royaume-Uni. Cette analyse révèle l'influence des modèles initiaux sur la réduction de l'écart lexical pour chaque ensemble de données, ainsi que l'effet du résumé sur la pertinence des termes sélectionnés.

   En outre, une comparaison entre notre système LoGE et deux systèmes proches de l'état de l'art en matière de récupération d'informations ad hoc non supervisées démontre la précision de notre méthode, tant en termes de justesse que de classement des résultats.

Ce chapitre est structuré de la manière suivante : la section 3.2 offre une étude exhaustive sur les travaux relatifs aux moteurs de recherche ad hoc, en se concentrant sur la vectorisation et l'expansion de texte. La section 3.3 expose en détail notre méthode et son intégration au sein du cadre LoGE est présentée dans la section 3.4. La section 3.5 aborde notre configuration expérimentale et analyse l'expansion des documents ainsi que son impact sur le processus de correspondance dans les moteurs de recherche ad hoc. Nous concluons dans la section 3.6.

3.2 État de l'art

Tandis que les techniques d'AP ont facilité l'élaboration de nouvelles méthodes en RI, elles comportent des limitations qui compliquent leur application pratique dans le secteur commercial. Un des obstacles majeurs de ces méthodes supervisées est la nécessité de disposer de données étiquetées. Ainsi, comme [Xio+17] certaines études partent du principe que des paires requête-document sont disponibles en abondance pour déterminer un score de correspondance entre elles. Un autre défi est le temps d'inférence, indépendamment de la durée importante d'apprentissage. Effectivement, pour une requête spécifique, il faut interroger le modèle sur l'ensemble du corpus documentaire pour identifier le document le plus pertinent à la requête. Même avec des ressources informatiques conséquentes, cela rend l'utilisation de ces modèles ardue et leur permet de produire des résultats dans un délai acceptable pour les utilisateurs. Notre approche vise à étendre les documents dans le but de réduire l'écart entre les requêtes et les documents, tout en préservant leurs pertinence, efficacité, adaptabilité à différents corpus documentaires et fiabilité sur des documents volumineux.

3.2.1 Recherche à partir de plongements vectoriels

La vectorisation est une approche permettant de traduire un mot ou un texte en un vecteur de valeurs numériques. Ces représentations permettent de s'adapter au contexte du corpus grâce à l'apprentissage des relations entre les mots basé sur les cooccurrences au sein du mêmes document, segment ou phrase. Dans la modélisation vectorielle, nous pouvons trouver différentes techniques : le Modèle Vectoriel [SWY75], le Plongement Lexical [Mik+13] et les Transformer/Modèle avec mécanisme d'auto attention [Vas+17]. Ces techniques ont facilité un traitement efficient des données en diminuant la dimensionnalité des documents écrits, ce qui a permis l'enregistrement de représentations simplifiées en vue de leur comparaison au sein d'un corpus. Leur première application a consisté à mesurer les occurrences de termes dans les documents, en les projetant dans un espace vectoriel commun [SWY75]. Cette projection de textes dans des espaces vectoriels a été considérablement influencée par le concept des Plongements Lexicaux [Mik+13], qui intègre le contexte de chaque mot dans la création de son vecteur, favorisant ainsi l'incorporation de la signification des mots.

BERT [Dev+19] (*Bidirectional Encoder Representations from Transformers* - Représentations de l'Encodeur Bidirectionnel issues des Transformeurs) prend en compte l'ensemble du contexte d'un mot

pendant la phase d'apprentissage. Un mot n'est pas limité à une représentation vectorielle statique ; il dispose plutôt d'un vecteur dynamique qui s'adapte selon le contexte particulier dans lequel il est employé. Cette approche permet de surmonter les limitations terminologiques inhérentes à d'autres modèles de Plongements Lexicaux, lesquels associent une représentation unique et invariable à chaque mot [Pfe+18 ; Nog+20 ; Zhe+20 ; Zhu+21 ; Wu+21b ; CD22 ; WLA22].

Toutefois, la représentation des documents demeure sous forme de vecteurs agrégés. Cette méthodologie impacte significativement la pertinence de l'information, qui risque de se focaliser sur une fraction restreinte du document global, submergeant ainsi les informations spécifiques portées par des mots distincts. Par ailleurs, les modèles fondés sur des réseaux neuronaux sont contraints par la dimension de leurs couches d'entrée. En conséquence, pour s'adapter à la structure du réseau, la taille des documents est réduite, entraînant l'exclusion de certaines informations [Kee+19 ; Zer+22 ; Li+23].

Néanmoins, les associations de méthodes vectorielles et symboliques font rarement l'objet d'études. Le système BISON [Sha+20] introduit une dimension supplémentaire d'information via un modèle d'attention, visant à attribuer des pondérations variées à chaque symbole en s'appuyant sur la mesure BM25 [RJ76].

Tout en conservant une représentation vectorielle lors du traitement des requêtes et des documents en ligne - et en engendrant certains des problèmes précédemment évoqués - cette méthode tire parti des avantages de ces deux approches : elle considère l'importance des mots au sein du corpus et des documents, intègre le contexte des mots et contribue à la réduction de la dimensionnalité du texte.

L'intégration des deux méthodologies apparaît comme la stratégie la plus performante pour élucider l'information sémantique des mots tout en préservant la transparence des classements. En conséquence, notre approche privilégiera non pas le domaine de la vectorisation, mais plutôt celui du symbolique, en exploitant des analyses statistiques pertinentes dans ce contexte.

3.2.2 Expansion de textes

L'objectif de l'enrichissement textuel est d'incorporer des informations supplémentaires au sein d'un texte, soit par le biais d'une reformulation, soit en y ajoutant des éléments additionnels, dans le but d'optimiser l'efficacité de la recherche. Dans le cadre de la recherche ad hoc, cette technique d'enrichissement peut être mise en œuvre tant pour les *requêtes* que pour les *documents*.

3.2.2.1 Expansion de requêtes

L'expansion de requête vise à maximiser le recoupement de la requête sur les documents. Ainsi, en proposant de nouveaux termes, la requête a plus de chances de correspondre. Le défi réside dans la gestion de la qualité de ces termes en fonction du corpus ciblé.

Dans le domaine de l'enrichissement de requêtes, la méthode du retour de pertinence ou « *Relevance feedback* » [RJ71] est particulièrement renommée. Elle autorise l'incorporation de termes supplémentaires à la requête originale. Ces termes proviennent des documents obtenus via un modèle de recherche plus conventionnel et sont choisis pour leur pertinence, souvent basée sur des caractéristiques statistiques, à l'instar du modèle RM3 [Jal+04]. L'utilisation de réseaux neuronaux pour combler l'écart lexical, en enrichissant la requête via les contextes des mots [Nas+21 ; Zhu+23], est également possible. Toutefois, bien que ces modèles affichent des performances prometteuses, leur efficacité dépend de la bonne fonctionnalité du moteur de recherche initial.

Les ontologies constituent des représentations structurées et explicites de la connaissance dans un domaine spécifique, établissant des liens clairs entre termes et concepts. Malgré une moindre vulnérabilité aux limites sémantiques, les relations étant formées en fonction du contexte et de la signification des mots, les ontologies restent des modules de traitement statiques, peu flexibles face à de nouveaux contextes. Cette rigidité est principalement due à la nécessité d'une expertise spécialisée pour leur élaboration, leur entretien et leur gestion [ATM14]. Un des défis majeurs de ces méthodes réside dans le fait qu'elles sont mises en œuvre en ligne, ce qui se traduit par une diminution des performances lors de la phase de correspondance du côté de l'utilisateur [Nog+19].

3.2.2.2 Expansion de documents

L'intérêt porté à l'enrichissement documentaire d'un corpus en RI est un phénomène récent. De nombreux travaux s'appuient sur l'étude de Billerbeck et Zobel (2005) [BZ05] pour étayer leur approche. Toutefois, cette étude se limite à l'emploi de méthodes basées sur la co-occurrence, sans prendre en compte les représentations sémantiques. En outre, l'argument de la rapidité mis en avant dans leurs travaux est pertinent principalement dans un cadre expérimental et ne se traduit pas nécessairement dans l'application pratique des algorithmes de recherche [Nog+19 ; Ma+22].

Certains travaux se concentrent sur la résolution de ces problèmes en utilisant l'apprentissage pro-

fond pour la RI en l'absence de paires requête/document. À titre d'exemple, le modèle UDEG [Jeo+21] se distingue. Ce modèle emploie un système de synthèse automatique de résumés pour traiter les documents [LC12], sa capacité à générer des abstractions lui permettant d'intégrer dans les résumés des mots absents des documents originaux. Un avantage notable de cette méthode est la possibilité de mener une recherche d'information de manière simple et rapide, comme avec BM25 [RJ76], une fois les résumés produits hors ligne. Cela démontre l'amélioration des performances permise par la réécriture des documents. Néanmoins, cette approche comporte des limites, notamment dans son applicabilité aux problématiques concrètes de domaines spécifiques. Bien qu'éliminant le besoin de paires requête/document alignées, elle s'appuie sur un système de générations de résumés, *Pegasus* [Zha+20a], qui requiert des paires résumé/document alignées pour son apprentissage. Ainsi, la nécessité des requêtes est indirectement substituée par celle des résumés, qui sont encore plus difficiles à produire manuellement, les deux approches étant par définition supervisées.

Un autre aspect limitatif de l'approche UDEG concerne son fondement sur le système *Pegasus*. Comme souligné auparavant, les réseaux de neurones se heurtent à des difficultés dans le traitement de documents volumineux. À titre illustratif, cette section dépasse largement la limite de taille de traitement de *Pegasus* et nécessiterait une réduction pour être traitée par le modèle, ce qui engendrerait une perte d'information significative et arbitraire. Par conséquent, cette contrainte diminue l'adaptabilité de la méthode à des jeux de données spécifiques.

Bien que les résultats obtenus soient prometteurs, nous sommes d'avis qu'une approche globale et non supervisée pourrait produire des résultats similaires, voire plus concluants, tout en surmontant les diverses difficultés évoquées. Dans cette optique, nous positionnons UDEG & PEGASUS (version adaptée à la RI) comme les méthodes les plus proches de l'état de l'art, en raison de leur aptitude à réécrire les documents de manière abstraite et de leur approche de traitement du texte hors ligne, privilégiant l'explicabilité des résultats.

3.3 Méthode d'expansion de documents

L'ambition de notre démarche est d'enrichir les documents avec des termes pertinents pour optimiser leur adéquation avec les requêtes. Cette approche se doit d'être non supervisée, en raison de la diversité des requêtes, tout en conservant une efficacité élevée sans nécessiter une réécriture onéreuse

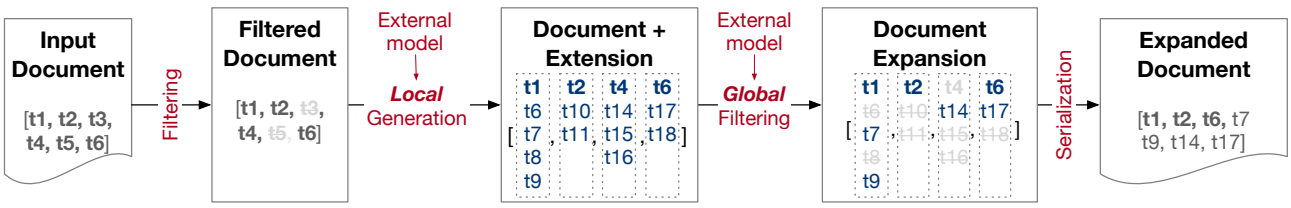

FIGURE 3.1 – LoGE - Un modèle Local et Global d'Expansion de documents

des requêtes. Notre objectif principal est donc d'améliorer le vocabulaire en s'appuyant sur des modèles pré-entraînés, afin de minimiser les discordances syntaxiques entre les requêtes et les documents dans le cadre de tâches de Recherche d'Information non supervisées. Nous exposons l'infrastructure de l'expansion documentaire et ses divers composants dans ce contexte.

3.3.1 LoGE - Un modèle Local et Global d'Expansion de documents

L'infrastructure d'expansion documentaire LoGE, que nous présentons dans la figure 3.1, est conçue pour optimiser la Recherche d'Information. Ce système renforce les moteurs de recherche en bonifiant le contenu textuel par l'intégration d'un contexte spécialisé qui fusionne une perspective à la fois globale et locale sur le contenu des documents.

1. Tout d'abord, chaque document soumis est traité par un processus de filtration, consistant à éliminer les termes non pertinents et les mots vides. Ce filtrage est basé sur un système de classement IDF traditionnel. Cette étape permet à notre méthode de se concentrer exclusivement et efficacement sur les termes qui apportent une particularité distincte au document, en regard du contenu global du corpus.
2. Par la suite, une phase de **Génération Locale** est mise en œuvre sur le document préalablement filtré. Cette phase vise à étendre le document en prévoyant l'ajout de termes, tout en considérant le contexte local de chaque terme. Cette opération s'appuie sur un modèle externe d'apprentissage automatique, qui peut être adapté selon les besoins (par exemple, BERT), et qui se base sur le contenu même du document pour générer des synonymes, effectuer des corrections orthographiques, entre autres. Les détails et discussions relatifs à cette étape spécifique sont exposés dans la section 3.3.3.
3. Lors de la génération de termes prédictifs pour l'extension du document, une phase de **Filtrage Global** est mise en place pour intégrer le contexte général associé au document. Cette phase

est cruciale pour se focaliser sur les termes ayant une importance accrue dans un domaine spécifique et, par conséquent, pour éliminer les termes qui n'apportent pas une valeur ajoutée pertinente. Cette étape peut être soutenue par divers modules. Concrètement, elle repose sur une représentation vectorielle des mots issue d'un modèle externe, projetant les mots dans un espace multidimensionnel (par exemple, SVD), et fonctionne comme un filtre en déterminant l'appartenance d'un terme à une dimension et/ou à un cluster spécifique.

4. Pour conclure, l'extension qui a subi le filtrage est intégrée et indexée dans le moteur de recherche sous la forme d'un **Document Étendu**.

Cette méthode offre plusieurs bénéfices intrinsèques :

1. C'est une *approche non supervisée*, caractérisée par l'absence de connaissance préalable concernant l'attribution spécifique des termes aux requêtes des utilisateurs et la détermination adéquate du document correspondant.
2. L'infrastructure proposée présente une capacité d'adaptation, autorisant l'échange de modèles d'apprentissage automatique tant locaux que globaux. Cette particularité offre la possibilité de calibrer finement le moteur de recherche en utilisant des modèles spécialisés ou universels, ce qui contribue à l'amélioration de la pertinence sur divers axes.
3. Pour terminer, la structure de `LoGE` se distingue par sa modularité, attribuable au fait que ses extensions sont générées en mode hors ligne. De ce fait, elle est conçue pour fonctionner de manière autonome, indépendamment de tout autre moteur de recherche.

L'objectif est l'ajustement du moteur de recherche au moyen de modèles d'apprentissage automatique, lesquels ne sont pas obligatoirement élaborés à partir du corpus sous-jacent, mais plutôt fondés sur des modèles prédéfinis externes (tels que BERT, LegalBERT, ClinicalBERT). Cette stratégie vise à atténuer le décalage lexical tout en préservant une vision à la fois locale et globale du corpus.

Exemple illustratif. L'application de notre méthode sera démontrée à l'aide d'un document constitué de six termes (de *t1* à *t6*), comme l'illustre la figure 3.1 (document initial). À chaque étape, ces termes seront manipulés afin de générer un document enrichi. La requête de l'utilisateur est définie comme (*t1, t14, t6*), où *t14* est un synonyme de *t4*, résultant ainsi en une augmentation de la superposition du vocabulaire entre la requête et le document.

3.3.2 Fonction de score

Dans le domaine de la recherche d'information, la fonction de score `BM25` [RJ76] est fréquemment employée dans les moteurs de recherche axés sur les termes. Elle présente un équilibre intéressant entre pertinence et efficacité. Pour les systèmes de recommandation non supervisés, une comparaison avec cette fonction et une tentative d'amélioration de son impact s'avèrent pertinentes.

Il convient de souligner que l'enrichissement du document ne modifie pas la fonction de score BM25, ce qui facilite son intégration dans divers moteurs de recherche. Cette synergie assure des avantages notables, tels que l'explicabilité (par la mise en exergue des termes), la modularité (permettant l'intégration dans les moteurs de recherche) et l'évolutivité (grâce à la distribution des documents et aux listes inversées).

Nous nous attachons par la suite à examiner l'effet de l'expansion du document sur cette fonction de score. Le modèle BM25 est défini comme suit :

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avg_{dl}})} \tag{3.1}$$

où $f(q_i, D)$ représente la fréquence du terme de requête q_i dans le document D. $IDF(q_i)$ représente la rareté du terme q_i au sein du corpus. $|D|$ est la taille du document, et avg_{dl} représente la taille moyenne des documents dans l'ensemble du corpus. $k_1 \in [1.0, 2.0]$ et $b = 0.75$ sont des paramètres libres.

3.3.3 Génération locale des extensions

L'étape de génération locale de notre méthode vise à aborder la problématique du manque de superposition de vocabulaire entre requête et document en augmentant le document avec une perspective locale. Pour chaque terme, elle génère un ensemble de synonymes ou de termes probables qui reflètent une proximité sémantique. Pour atteindre cet objectif, un *Modèle Externe* est nécessaire, capable de fournir une analyse locale des termes (par exemple, une fenêtre sur le document) afin de formuler des recommandations de termes. Ces recommandations sont alignées avec les termes les plus probables en rapport avec la séquence des termes et un modèle pré-entraîné.

Le modèle reposant sur le masquage de mot, BERT [Dev+19], est idéalement adapté à nos be-

soins, car il évalue la probabilité d'apparition d'un terme masqué dans une phrase ou un passage. Word2Vec [Mik+13 ; Chu17] pourrait aussi être adapté pour recommander des termes en suggérant des synonymes pour un terme spécifié dans une phrase. Ces modèles partagent une propriété fondamentale sur laquelle repose notre proposition. Ils permettent de générer des termes probables dans un domaine de spécialisation donné sans nécessiter d'entraînement ad hoc sur notre corpus de recherche.

De plus, étant donné que les requêtes des utilisateurs peuvent présenter un vocabulaire légèrement différent de celui du corpus, un modèle spécifique à un domaine, plutôt que spécifique au corpus, permettra d'élargir le vocabulaire. L'impact de divers modèles spécifiques à un domaine sur cette étape sera examiné dans l'étude expérimentale détaillée dans la section 3.5.

Comme indiqué précédemment, le modèle d'entrée requiert une séquence de termes afin de prédire les termes les plus probables, en tenant compte de la vue locale (c'est-à-dire, de la séquence). Il est important de noter que le modèle d'entrée est contraint par la taille de la séquence pour effectuer des prédictions. Cela implique un processus de fenêtrage sur la séquence globale (c.-à-d. le document) pour extraire des sous-séquences.

Diverses stratégies de fenêtrage peuvent être envisagées, allant de simples fenêtres glissantes à la segmentation en phrases. Dans ce contexte, les phrases sont traitées comme des entrées du module de *fenêtrage*.

3.3.3.1 Fenêtrage de documents

À partir de la séquence de termes $\mathcal{S}$ correspondant au document d'entrée filtré D, et d'un modèle de prédiction d'entrée $\mathcal{B}$ avec un seuil de limitation σ, nous appliquons une fenêtre sur le texte pour extraire des séquences de termes $\mathcal{S}_i \subseteq \mathcal{S}$ telles que $|\mathcal{S}_i| \leq \sigma$. Cette étape est illustrée dans la deuxième étape de la figure 3.2 où $\sigma = 4$, ce qui produit 3 fenêtres distinctes ; la deuxième n'est pas utilisée puisque $t3$ est un terme filtré et ne produit pas de prédictions.

Ensuite, les $\mathcal{S}_i$ obtenus sont utilisés comme contextes pour le modèle de prédiction. $\mathcal{B}$ prédit pour un terme masqué $t_j \in \mathcal{S}_i$ un ensemble de termes pertinents définis comme t_j^k avec une probabilité p_k associée tel que :

$$\mathcal{B}(S_i, j) = [(t_j^1, p_1), ..., (t_j^k, p_k)]$$

Cette étape est illustrée dans la troisième et quatrième étape de la figure 3.2, avec 3 fenêtres et des termes masqués qui génèrent des prédictions créant l'extension du document.

Afin d'appliquer la prédiction pour chaque terme masqué, il est nécessaire d'itérer sur les séquences S_i. Pour préserver le contexte, le terme masqué est positionné au centre de la séquence, mais pour les séquences S_0 et $S_{|S|-1}$, il est nécessaire de prédire respectivement les termes de début et de fin de la séquence. La formule suivante formalise le domaine de valeurs pour les termes masqués (position j) par la fonction de fenêtrage $\mathcal{W}$:

$$\mathcal{W}(S_i) = \begin{cases} j \in \left[0, \left\lfloor \frac{|\mathcal{S}_i|}{2} \right\rfloor \right[& \text{si } i = 0, \\ j \in \left[\left\lfloor \frac{|\mathcal{S}_i|}{2} \right\rfloor, |\mathcal{S}_i| \right[& \text{si } i = |\mathcal{S}| - 1, \\ j = \left\lfloor \frac{|\mathcal{S}_i|}{2} \right\rfloor - 1 & \text{sinon} \end{cases}$$

Alors, l'extension T_{ext} générée peut être formalisée par l'équation 3.2. Pour chaque terme t_j de la séquence S, les prédictions $\mathcal{B}(S_i, j)$ sont générées à la position j selon la fenêtre $\mathcal{W}(S_i)$. Les prédictions sont produites uniquement si t_j n'a pas été préalablement filtré (par la fonction *filtré(S)*) lors de la première étape du traitement.

$$t_{ext} \in T_{ext} \quad | \quad \forall i \in [0, |S|[, j \in \mathcal{W}(S_i), t_{ext} \in \mathcal{B}(S_i, j) \wedge t_j \notin \text{filtré}(S) \tag{3.2}$$

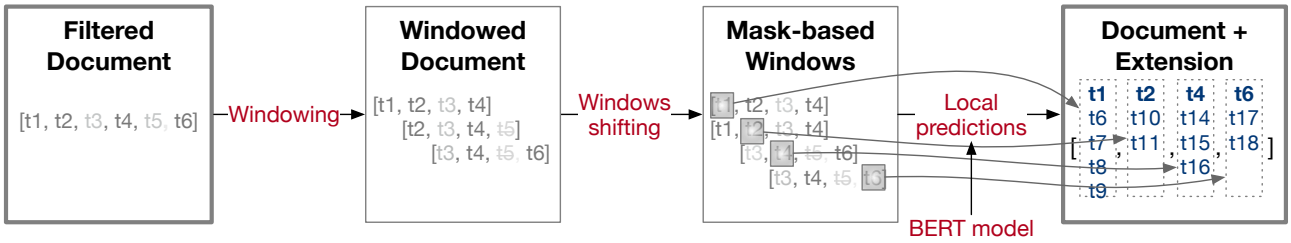

FIGURE 3.2 – Étape de génération de termes par fenêtrage local via BERT

Les quatre prédictions représentées au sein de la troisième étape de la figure 3.2 illustrent le décalage de t_j sur la fenêtre correspondante : $j \in [0, 1]$ pour la fenêtre initiale, $j = 1$ pour la fenêtre intermédiaire et $j \in [1, 3]$ pour la dernière fenêtre, à l'exception du terme t_5 ($j = 2$ dans la dernière fenêtre), car il a été filtré. Néanmoins il est à noter que le terme demeure au sein de la fenêtre, car le modèle $\mathcal{B}$ nécessite la séquence de termes complète afin d'effectuer une prédiction dans ce contexte.

3.3.3.2 Modèle de prédiction externe interchangeable

Comme le modèle de prédiction externe n'est pas nécessairement entraîné sur le corpus sous-jacent, il peut prédire des termes relativement au domaine sur lequel il a été entraîné. Il peut être général comme `BERT` ou spécialisé comme `LegalBERT`. Ce choix aura un impact certain sur la qualité des prédictions. De plus, un modèle externe pré-entraîné peut proposer des termes initialement absents du corpus comme des synonymes améliorant la qualité de la superposition du vocabulaire entre documents et requêtes.

3.3.4 Impact de l'extension du document sur le score

La génération d'une extension entraîne l'élargissement du document de sortie, impactant ainsi la correspondance et l'évaluation des scores. Le but premier est d'enrichir le document avec un nombre accru de termes, dans le but d'accroître la probabilité de correspondance avec les requêtes. Toutefois, cet enrichissement entraîne un effet secondaire : l'introduction de bruit dans le document, ce qui peut affecter négativement le score.

L'influence des termes q_i sur le score défini dans l'équation 3.1 peut être exprimée comme suit (pour des raisons de simplification, nous partons du principe que $k_1 \sim 1 + \epsilon$ et que $b \sim 1 - \epsilon$) :

$$\begin{aligned} impact(q_i) &= IDF(q_i) \cdot \frac{2 \cdot f(q_i, D)}{f(q_i, D) + (2 + \epsilon)\left(\epsilon + (1 - \epsilon)\frac{|D|}{avg_{dl}}\right)} \\ \frac{1}{impact(q_i)} &= \frac{f(q_i, D) + 2 \cdot \frac{|D|}{avg_{dl}}}{2 \cdot f(q_i, D) \cdot IDF(q_i)} \\ &= \frac{f(q_i, D)}{2 \cdot f(q_i, D) \cdot IDF(q_i)} + \frac{2 \cdot \frac{|D|}{avg_{dl}}}{2 \cdot f(q_i, D) \cdot IDF(q_i)} \\ &= \frac{1}{2 \cdot IDF(q_i)} + \frac{\frac{|D|}{avg_{dl}}}{f(q_i, D) \cdot IDF(q_i)} \end{aligned} \tag{3.3}$$

En tenant compte du fait que la création de l'extension T_{ext} n'affecte pas l'IDF et qu'elle exerce une influence uniforme sur la taille moyenne des documents, l'effet positif des termes q_i inclus dans T_{ext}, tel qu'énoncé dans l'équation 3.3, peut être résumé de la manière suivante :

$$impact(q_i) \sim \frac{f(q_i, D)}{|D|} \tag{3.4}$$

Maintenant, en considérant un impact positif lors de l'ajout d'une extension au document $|D| + |T_{ext}|$, l'impact d'un terme q_i dans l'équation 3.4 est positif si :

$$\begin{aligned}
\frac{f(q_i, D)}{|D|} &< \frac{f(q_i, D + T_{ext})}{|D| + |T_{ext}|} \\
\frac{f(q_i, D)}{|D|} &< \frac{f(q_i, D) + f(q_i, T_{ext})}{|D| + |T_{ext}|} \\
\frac{f(q_i, D) \times (|D| + |T_{ext}|)}{|D| \times (|D| + |T_{ext}|)} &< \frac{(f(q_i, D) + f(q_i, T_{ext})) \times |D|}{|D| \times (|D| + |T_{ext}|)} \\
\frac{f(q_i, D) \times (|T_{ext}|))}{|D| \times (|D| + |T_{ext}|)} &< \frac{f(q_i, T_{ext}) \times |D|}{|D| \times (|D| + |T_{ext}|)} \\
f(q_i, D) \times \frac{|T_{ext}|}{|D|} &< f(q_i, T_{ext})
\end{aligned} \tag{3.5}$$

Selon les conclusions tirées de l'équation 3.5, il est possible d'affirmer qu'un terme inclus dans l'extension et correspondant à la requête produit un effet positif si :

1. Le terme q_i est nouveau. Toujours vrai, ce qui est l'objectif de notre étape d'extension de terme.
2. Si le terme q_i était préalablement présent dans D, son inclusion augmente la pertinence du terme qui se trouve dans le document ainsi que dans l'extension.

Ceci implique que le processus de génération locale doit identifier non seulement des termes similaires, mais aussi valoriser ceux qui étaient pertinents dans le document antérieur. Néanmoins, il est important de souligner, comme le montre l'équation 3.5, que la dimension de $|T_{ext}|$ exerce une influence significative sur l'ensemble des termes de l'extension. En effet, cette dimension peut se révéler préjudiciable si :

1. Si le terme q_i est présent dans D mais absent de T_{ext}, son influence décline proportionnellement à l'augmentation de la taille de l'extension.
2. Les occurrences de q_i dans T_{ext} doivent avoir une contribution supérieure au rapport $\frac{|T_{ext}|}{|D|}$ (autrement dit, $\frac{|T_{ext}|}{k_1 \cdot |D|}$, $k_1 = 1$ dans le scénario le moins favorable).

Il est possible de déduire que l'incorporation de termes supplémentaires dans le document influence le niveau de bruit. Il est nécessaire de limiter la taille de $|T_{ext}|$ afin d'assurer un impact moyen positif des termes issus de l'extension. Par conséquent, quand $|T_{ext}| \leq |D|$, chaque terme ajouté présente un effet bénéfique, y compris pour $q_i \in D$, et pour $q_i \notin T_{ext}$, l'impact est au plus réduit de moitié.

Ainsi, l'optimisation de l'expansion du document peut être conceptualisée par la maximisation des probabilités des termes introduits durant le processus de génération locale. Cette question peut être formulée de la manière suivante :

$$\underset{q_i \in T_{ext}}{\arg\max} \left(\frac{f(q_i)}{|T_{ext}|} \right) = \begin{cases} q_i | & f(q_i) = \sum_{i=1}^{n \leq |D|} f(q_i, T_{ext}) \\ & q_i \in \mathcal{B}(D) \end{cases} \tag{3.6}$$

En conclusion, l'optimisation de l'extension correspond à la maximisation d'un ensemble de prédictions générées par $\mathcal{B}(D)$. Différentes stratégies de sélection peuvent être mises en œuvre pour affiner ce choix, contribuant ainsi à la diminution du bruit. Cependant, bien que la prédiction locale soit efficace pour identifier des termes pertinents spécifiques au document, elle ne garantit pas nécessairement leur pertinence dans une perspective globale du corpus. Le poids d'un terme dans le cadre d'une requête spécifique doit être considérablement filtré dans un contexte global plutôt que local. Par conséquent, la procédure de filtrage lors de la création de l'extension (autrement dit, *argmax*) doit s'inscrire dans une approche globale du corpus, tel qu'il sera exposé dans les développements suivants.

3.3.5 Filtrage global des extensions

L'accroissement de la taille de l'extension est proportionnel à celle du document. En conséquence, la pertinence des termes diminue en raison de l'ajout de termes superflus au document. Les termes incorporés lors de l'étape antérieure sont spécifiques au document, reflétant une perspective locale. De ce fait, le but du filtrage de l'extension est de minimiser le nombre de termes, autrement dit, de réduire le bruit, en représentant le document tout en intégrant une conception globale pour déterminer la pertinence ou l'irrégularité des termes.

Pour réaliser cela, il est nécessaire d'extraire les termes clés d'un document donné en se basant sur leur représentation dans un contexte global. Cette démarche implique l'utilisation d'un plongement de mots externe et d'une représentation du document par *Décomposition en Valeurs Singulières* (SVD).

Le plongement lexical fournit pour chaque terme du document et de son extension une représentation spécifique au domaine concerné. L'objectif est d'employer un modèle adapté au domaine pour saisir la dimension sémantique dans une perspective globale.

L'étape de *SVD* offre une représentation synthétique du document en calculant l'ensemble de ses

plongements et en déterminant les axes les plus représentatifs. Ces axes peuvent être interprétés comme des thèmes ou sujets émanant du *nouveau* document.

En conclusion, le document et son extension sont filtrés en se fondant sur ces dimensions. Le filtrage peut consister à retenir les termes associés à ces axes (caractéristiques), éloignés d'eux (plus distinctifs), ou appartenant à un groupe spécifique, entre autres. L'objectif final est de générer une liste de termes dont le « *bruit* » est minimal. En effet, comme vu dans l'équation 3.6 les mots pertinents ont un meilleur impact si la longueur de l'extension est minimale. Ainsi, un trop grand nombre de mots rajoute de l'information non significative (les termes superflus ne contribuant pas à la représentation principale du document) et ainsi génère du bruit. En se focalisant sur les dimensions propres au document, cette liste de mot devient pertinente, réduit le *bruit* généré et augmente l'impact de ces termes.

De plus, sans cette étape de filtrage, les mots de l'extension risque d'être présents dans de nombreux documents et les rends plus proches les uns des autres et donne un IDF quasi nul. La pertinence est donc réduite par la proximité plus grande avec les autres documents du corpus « *étendu* ».

La sélection de termes significatifs peut être réalisée par voisinage au moyen d'un algorithme des k-plus proches voisins (kNN) par exemple ou sur la base d'un seuil prédéfini. Chacune de ces méthodes influence la qualité du filtrage. Néanmoins, le kNN assure une maîtrise de la taille de l'extension, tandis que l'utilisation d'un seuil permet de retenir l'ensemble des termes pertinents.

Par conséquent, la procédure de filtrage global peut être formalisée comme suit, conformément à l'équation 3.7. Un terme t_{exp} est incorporé dans l'expansion s'il est présent dans l'extension T_{ext}, et si son plongement $\mathcal{E}(t_{ext})$ contribue à la réduction du bruit en fonction de la fonction de distance $\mathcal{D}$. Dans ce contexte, le bruit est défini par la distance par rapport aux axes de la Décomposition en Valeurs Singulières (SVD_d), laquelle doit être inférieure à un seuil σ.

$$t_{exp} \in T_{exp} \quad | \quad t_{exp} \in T_{ext} \wedge \mathcal{D}\left(\mathcal{E}(t_{ext}), SVD\right) < \sigma \tag{3.7}$$

La figure 3.3 illustre la procédure de filtrage global. Initialement, un plongement lexical externe est employé pour la vectorisation du document, incluant l'ensemble des termes générés lors de *l'étape de génération locale*. Par la suite, une représentation par SVD est appliquée au document. Chaque terme est ensuite projeté sur le SVD et retenu s'il contribue à minimiser la distance avec les axes.

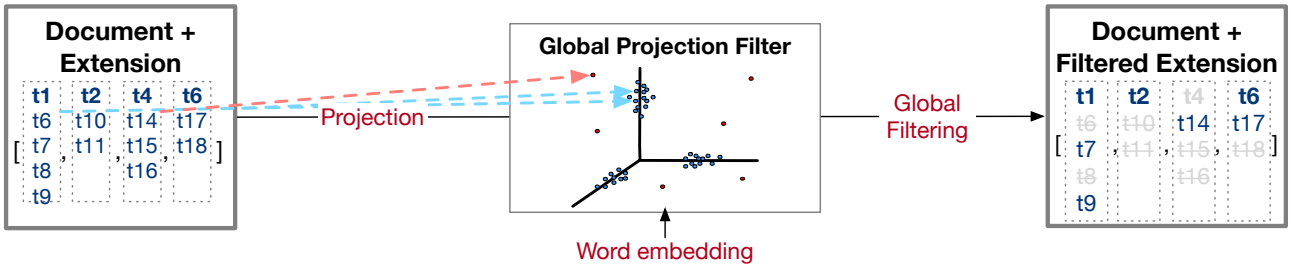

FIGURE 3.3 – Filtrage global des termes proposés

Il est pertinent de souligner que les termes initiaux peuvent être exclus par le SVD s'ils sont jugés « *excessivement bruyants* » (par exemple, le terme *t4*).

Il est à noter que cette étape de filtrage peut être adaptée en utilisant une variété de plongements lexicaux et en appliquant différents filtres pour affiner cette perspective globale sur les documents.

La formulation du score d'un mot ω est définie comme suit :

$$Score(\omega) = \frac{1}{n} \sum_{i=1}^{n} \frac{\overrightarrow{\omega} \cdot \overrightarrow{a_i}}{\| \overrightarrow{\omega} \|} \tag{3.8}$$

où un vecteur de plongement lexical $\overrightarrow{\omega}$ est jugé davantage représentatif lorsqu'il se trouve à proximité des n axes de la SVD $\overrightarrow{a}$.

3.4 Architecture de LoGE

L'approche LoGE que nous avons développée se caractérise par sa modularité. Le processus d'expansion des documents, opérant hors ligne et dans le domaine symbolique, s'appuie sur des modèles interchangeables selon les besoins.

Dans cette optique, nous avons élaboré un cadre modulable intégrable aux moteurs de recherche standards, comme Elasticsearch [1], pour augmenter la portée de notre système [Zhu+22]. Dans ce contexte, une architecture spécifique a été mise en place (c.f. figure 3.4), permettant la génération hors ligne de l'expansion des documents via des processus indépendants, tout en assurant la continuité opérationnelle du système.

L'efficacité du système repose essentiellement sur le traitement hors ligne, qui consiste à ajouter

1. https//elastic.co/

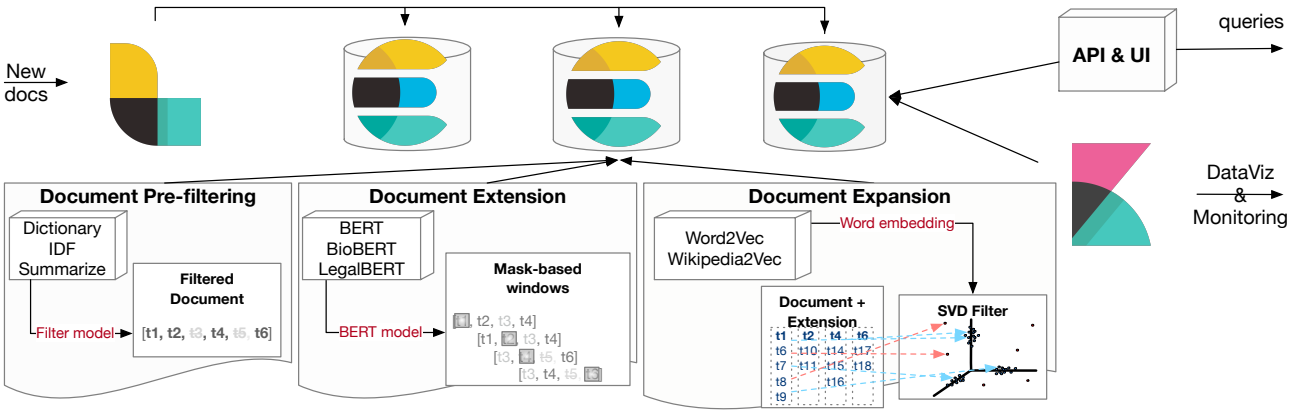

FIGURE 3.4 – Architecture de LoGE

des termes aux documents préalablement à leur interrogation. La capacité d'extension du système est naturellement gérée par le moteur de recherche, qui utilise des listes inversées et assure un passage à l'échelle linéaire grâce à la distribution sur un cluster, ici composé de trois instances.

3.4.1 *Stack* ELK

L'ensemble de la suite ELK est employé pour l'importation des documents via `LogStash` directement au sein du cluster de nœuds Elasticsearch, et pour la création d'un tableau de bord de suivi via `Kibana`. Les textes sont archivés sous la forme de documents JSON, lesquels sont par la suite étendus de manière autonome à travers des déclencheurs, comme démontré précédemment (voir l'exemple dans la figure 3.5). Dans la phase initiale, le document se compose uniquement du texte source avec les termes (t1, t2, ..., t6), représentés par la clé « *text* ».

3.4.2 Pré filtrage de document

Le service initial est chargé de surveiller le moteur de recherche pour détecter les documents entrants. Chaque fois qu'un document doit subir un filtrage, consistant à éliminer les termes non pertinents et les mots vides (en l'absence d'une clé « *filter* », comme le montre la requête DSL Elasticsearch ci-dessous), ce service procède au traitement du document. Il applique ladite étape de filtrage et actualise le document avec une liste renouvelée de termes (t1, t2, t4, t6).

```
{"bool": {
 "must_not": [{"exists": {"field": "filter"}}]
```

```
{ "id" : 1,
  "text" : "t1 t2 t3 t4 t5 t6",
  "filter" : ["t1","t2","t4","t6"],
  "extension":{ "LegalBERT":[
    [ {"term":"t1","score":0.9},
      {"term":"t6","score":0.83},
      {"term":"t7","score":0.7}],
    [ {"term":"t2","score":0.7},
      {"term":"t10","score":0.81},
      {"term":"t11","score":0.83}],
    [ {"term":"t4","score":0.75},
      {"term":"t14","score":0.81},
      {"term":"t15","score":0.83}],
    [ {"term":"t6","score":0.6},
      {"term":"t17","score":0.61},
      {"term":"t18","score":0.3}]
  ]},
  "expansion" : {
    "LegalBERT_summVD":[
      [ {"term":"t1","score":0.9},
        {"term":"t7","score":0.7}],
      [ {"term":"t2","score":0.7}],
      [ {"term":"t14","score":0.81}],
      [ {"term":"t6","score":0.6},
        {"term":"t17","score":0.61}]
    ]
  }
}
```


FIGURE 3.5 – Exemple de document avec les termes initiaux, les termes filtrés, les extensions générées avec `LegalBERT` et ces expansions avec SummVD.

}}

L'étape de filtrage du document original est essentielle pour diminuer le nombre de tokens vides et sans pertinence, pouvant influencer les résultats obtenus. Les filtres appliqués sont standardisés et incluent la ponctuation, les éléments numériques, les mots vides et les caractères isolés.

Par ailleurs, les termes fréquemment rencontrés dans le corpus global sont jugés peu significatifs pour le document spécifique. Par conséquent, la génération d'extensions pour ces termes, susceptibles de produire des synonymes communs à de nombreux documents, est évitée. Ainsi, la dimension de la clé « *filter* » est régulée par cette procédure.

3.4.3 Extension de documents

Pour les documents possédant une clé « *filter* » mais pas de clé « *extension* » associée au modèle de prédiction spécifié, le service en question, similaire au précédent, procède à la génération de l'extension. Ce service se base sur un modèle BERT pour enrichir le document au moyen de prédictions terminologiques. Comme détaillé dans la section 3.3.3.1, des fenêtres glissantes sont appliquées à la liste des termes sous la clé « filter ».

Les listes de prédictions sont ensuite enregistrées sous une nouvelle clé « *extension* », en parallèle du modèle utilisé (à savoir, *LegalBERT*). Les termes prédits et leurs scores sont archivés sous forme de tableaux pour chaque terme filtré. Ce mode de stockage assure à la fois l'indépendance des services, la variabilité des processus de post-filtrage, la possibilité de formuler différentes requêtes, et l'explicabilité (grâce à la traçabilité).

Concernant les modèles d'extension, nous optons exclusivement pour des modèles de type remplissage de masque (notamment BERT), qui masquent un terme du document et prédisent un ensemble d'éléments probables. Pour évaluer l'impact du modèle sélectionné sur des ensembles de données spécialisés, comme indiqué dans la section 3.5, nous avons choisi un modèle général ainsi que des modèles spécifiques aux domaines concernés.

Dans le but de conserver une approche symbolique centrée sur les termes, le modèle BERT [Dev+19] a été adapté pour éviter la génération de syllabes et pour maintenir les termes aux vecteurs inconnus dans leur forme originale, garantissant ainsi la production de termes compréhensibles en sortie. Cette adaptation a permis d'observer une amélioration notable des résultats et une diminution significative de la discordance du vocabulaire.

BERT [2] se positionne comme un modèle général pré-entraîné de référence, riche d'une connaissance étendue acquise à partir d'une vaste collection de données couvrant de multiples domaines.

BioClinicalBERT [Als+19], intensivement formé sur des articles de *PubMed* et des notes médicales, a été utilisé en veillant à ce que les documents présents dans notre corpus ne fassent pas partie de ses documents d'entraînement, afin d'éviter tout avantage en matière de connaissance spécifique aux documents (p. ex. à se ramener à un cadre supervisé).

2. `https://huggingface.co/bert-base-uncased`

LegalBERT [Cha+20], modèle spécifique au domaine juridique, se distingue par un corpus de pré-entraînement diversifié, incluant des textes de législation de l'UE et du Royaume-Uni, ainsi que des jurisprudences et contrats juridiques. Comme pour *BioClinicalBert*, les documents de notre corpus juridique sont soigneusement écartés en cas de présence dans ses documents d'entraînement.

3.4.4 Expansion de documents

Pour produire l'expansion du document, chaque extension requiert un filtrage. Des stratégies variées peuvent être appliquées sur les extensions et enregistrées sous forme de nouvelles clés « *expansion* ».

Concernant le **filtre Top-k**, la quantité de termes prédits pour chaque jeton masqué de BERT pose problème. Nous avons observé que la probabilité de chaque terme figurant dans la liste générée par les modèles de langage susmentionnés diminue considérablement et rapidement. Par conséquent, nous avons établi un seuil de 10 termes principaux pour chaque prédiction afin de réduire le bruit et le temps de calcul des expansions générées.

Quant au **filtre de Fréquence**, malgré l'indépendance de la requête dans l'expansion du document (p. ex. une génération non supervisée), il est nécessaire de minimiser certaines redondances pour améliorer la qualité de l'extension. Ainsi, les termes sont traités en fonction de l'ensemble du corpus pour limiter les répétitions induites par BERT, pouvant indiquer un entraînement inadéquat ou un corpus de base inapproprié sur des modèles pré-entraînés. Les extensions sont donc filtrées en fonction du taux d'apparition des mots générés par rapport à l'extension complète de chaque document.

Le **filtre SummVD** représente la dernière étape de post-traitement de notre méthode. Il s'agit d'une adaptation de la chaîne de traitement de SummVD [She+22] décrit au chapitre précédent, visant à produire un résumé extractif non supervisé basé sur des thèmes identifiés par clustering des vecteurs de termes de documents, utilisant la SVD pour déterminer les dimensions les plus significatives. Les vecteurs thématiques sont comparés à tous les mots du document pour calculer une similarité cosinus avec chaque thème. Une phrase du texte original est alors choisie pour chaque thème en cumulant les scores de ses termes.

Dans notre contexte de RI, notre apport se concentre sur la phase de sélection des termes. Travaillant au niveau des mots plutôt que des phrases, nous retenons les termes qui maximisent la somme de leurs similarités avec chaque thème, en assignant des poids conformes au classement établi par la

SVD.

3.4.5 Module de recherche

Pour terminer, l'infrastructure LoGE exécute des requêtes Elasticsearch paramétrées selon l'extension générée (dans ce cas, "*LegalBERT_summVD*"), comme illustrée dans la requête ci-dessus. Lors de l'application de modèles BERT ou de filtres différents, la requête peut être ajustée pour tirer parti de l'extension adéquate, illustrant ainsi la modularité de notre méthode. De plus, cette approche met l'accent sur la mise en évidence des termes correspondants dans le but d'améliorer l'explicabilité des résultats obtenus.

```
{
    "bool": {
        "should": [
            {
                "more_like_this": {
                    "fields": ["extension.LegalBERT_summVD"],
                    "like": "t1 t7 t17"
                }
            }
        ]
    },
    "highlight":{"fields":{"extension.LegalBERT_summVD":{}}}
}
```

Pour appliquer la fonction de score **BM25** afin de classer les documents pour chaque recherche, nous avons utilisé la requête suivante. Elle doit être définie lors de la mise en place de l'architecture qui précède l'importation du corpus. En fait, cette méthode s'est avérée la plus adéquate au fil des décennies et surpasse toute concurrence lorsqu'il s'agit de faire correspondre et de noter des termes dans une recherche ad hoc.

```
{
    "settings": {
```

```
        "number_of_shards": 3,
        "index" : {
            "similarity" : {
                "default" : {
                    "type" : "BM25",
                    "b": 0.75,
                    "k1": 1.2
                }
            }
        }
    }
}
```

L'intégralité de l'architecture (disponible sur Github [3]) a été conçue sous la forme d'un cluster Docker, comprenant divers services qui offrent la possibilité d'ajuster le nombre de noeuds *Elastic*, ainsi que la quantité et le type de modèles et de filtres.

3.5 Étude expérimentale

Nous abordons à présent l'étude expérimentale de la méthode que nous proposons. Tout d'abord, l'analyse se concentre sur l'effet des modèles spécialisés dans la génération d'extensions, avec pour objectif d'évaluer leur adaptabilité.

Par la suite, une investigation approfondie est menée sur l'amélioration des extensions, en les comparant aux modèles non supervisés issus des travaux connexes, afin de résoudre une recherche ad hoc.

3.5.1 Jeux de données

Pour la réalisation de nos expériences, nous avons recours à deux ensembles de données spécialisés, relevant de domaines complexes qui exigent une expertise terminologique spécifique. Les corpus choisis comprennent des textes d'une longueur supérieure à celle habituellement observée dans les travaux de

3. Github sourcecode : `https://github.com/leonard-de-vinci/LoGE_DocExt_BERT-FILTER` (GNU GPL v2)

TABLE 3.1 – Caractéristiques des corpus NFCorpus & EU2UK

	Quantile	NF Corpus	EU2UK
Nombre de tokens moyens par document		**4 766,33**	**11 447,98**
Quantiles des tokens par document	0,25	**3 557,5**	**316,0**
	0,5	**3 940,0**	**1 207,0**
	0,75	**5 634,0**	**7 106,5**
Nombre de tokens moyens par requête		**1 904,94**	**7 314,7**
Quantiles des tokens par requête	0,25	**42,0**	**1 126,25**
	0,5	**237,0**	**2 532,5**
	0,75	**2 098,0**	**6 220,75**
Nombre moyen de relations par requête		**1,76**	**1,04**
Quantiles des relations par requête	0,25	**1,0**	**1,0**
	0,5	**1,0**	**1,0**
	0,75	**2,0**	**1,0**

référence actuels, s'étendant de quelques centaines de termes à plusieurs milliers dans nos cas d'étude.

3.5.1.1 NFCorpus

NFCorpus [Bot+16] est un ensemble de données médicales en langue anglaise reconnu dans le domaine de la Recherche d'Informations, présente des caractéristiques distinctives.

Selon les données du tableau 3.1, on observe une distribution hétérogène de la longueur des requêtes, variant de quelques mots à plusieurs milliers, bien que cette longueur reste inférieure à celle des documents. Le nombre moyen de documents correspondant à chaque requête est relativement bas, soulignant ainsi l'importance d'affiner la précision dans la sélection des documents.

3.5.1.2 EU/UK legislation

EU/UK legislation [Cha+21] est un ensemble de données nouvellement publié en anglais, centré sur le traitement de longs documents/requêtes juridiques. Ce corpus est principalement orienté vers l'appariement de textes législatifs étendus et similaires, avec une variabilité plus marquée en matière de taille de document plutôt que de requêtes.

Par conséquent, il est possible d'identifier des caractéristiques distinctives, étant donné que la correspondance entre les requêtes et les documents est presque exclusivement unique, avec une longueur substantielle pour les deux catégories. Ce corpus se distingue également par une longueur moyenne de requête considérablement élevée par rapport à la longueur des textes.

TABLE 3.2 – Évolution des résultats de recherche utilisant uniquement des extensions générées par un modèle de langage dans `LoGE` et filtrés avec différents seuils minimums de probabilité des mots générés (excluant tout autre processus de filtrage des expansions).

	Modèle	σ	MAP @10	F1-score @10	NDCG @10	Prec. @10	Rappel @10	Taille des extensions	Chevauchement de vocabulaire
EU2UK	BERT	0	0,1044	0,0876	0,3712	0,0485	0,4681	32 913,28	**0,4367**
	BERT	0,10	0,1288	0,0925	0,4383	0,0515	0,4877	4 453,44	0,2281
	BERT	0,40	0,1301	0,1032	0,4564	0,0574	0,5466	2 467,00	0,1467
	LEGAL	0	0,1291	0,0981	0,4465	0,0544	0,5196	34 228,94	0,4159
	LEGAL	0,10	0,1294	0,0977	0,4370	0,0544	0,5098	4 960,30	0,2219
	LEGAL	0,40	**0,1400**	**0,1137**	**0,4851**	**0,0632**	**0,5980**	3 343,18	0,1670
NFCorpus	BERT	0	0,0892	0,1143	0,3168	0,0693	0,4669	14 994,48	**0,5404**
	BERT	0,10	**0,1077**	**0,1235**	**0,3699**	**0,0752**	**0,5115**	1 893,26	0,2784
	BERT	0,40	0,1029	0,1195	0,3566	0,0731	0,4888	737,36	0,1644
	BIO	0	0,0626	0,0860	0,2446	0,0534	0,3413	10 939,04	0,1412
	BIO	0,10	0,0441	0,0611	0,1639	0,0390	0,2223	231,26	0,0186
	BIO	0,40	0,0536	0,0586	0,1846	0,0377	0,2136	13,82	0,0020

3.5.2 Enrichissement et adaptabilité

L'utilisation de modèles pré-entraînés pour la génération d'extensions diminue la discordance de vocabulaire entre les requêtes et les documents associés, en comblant les lacunes des synonymes de termes dans un contexte spécifique, c'est-à-dire en matière de terminologie. Comme illustré au tableau 3.2, le chevauchement des termes uniques dans les requêtes sur des documents préalablement filtrés s'accroît significativement pour les deux ensembles de données (à un seuil de 0). Toutefois, l'augmentation considérable de la longueur des extensions a un impact négatif sur les résultats de recherche, particulièrement lorsque la taille du document initial est très élevée.

Le filtrage des probabilités de mots générées par les modèles reposant sur BERT améliore les performances (0,4 pour LegalBert, et 0,1 pour BERT sur NFCorpus). Le chevauchement de vocabulaire représente le pourcentage moyen de mots en commun entre requête et documents. Bien que le chevauchement de vocabulaire se réduise, les termes sélectionnés affichent des performances satisfaisantes dans la majorité des cas, contribuant ainsi à la réduction du bruit. Par conséquent, il est nécessaire d'équilibrer un meilleur chevauchement avec l'application de termes spécifiques, impliquant une étape de filtrage supplémentaire après les prédictions.

L'écart lexical reflète la spécificité de chaque ensemble de données. Par exemple, EU2UK souffre d'un manque de termes spécifiques que LegalBert pourrait apporter, car avec un seuil de 0,4, les termes pertinents sont des synonymes spécifiques. En revanche, NFCorpus contient des requêtes avec

des termes spécifiques adéquats, mais ceux-ci ne correspondent pas aux termes communs. Ainsi, le modèle BERT a un effet positif, mais nécessite un seuil de 0,1 pour atténuer le bruit.

L'adaptabilité de notre méthode réside dans l'emploi de modèles de langage pré-entraînés, qu'ils soient généralistes ou spécialisés, sans nécessiter de formation supplémentaire. Cette caractéristique facilite son déploiement dans un contexte professionnel et réduit le temps requis pour l'installation de l'architecture.

3.5.3 Impact du filtrage de l'extension

Désormais, nous nous concentrons davantage sur la pertinence que sur le chevauchement afin de réduire le bruit produit dans les extensions. Cette phase englobe la synthèse de l'extension élaborée. Comme développé à la section 3.3.5, `LoGE` génère des expansions de documents à partir de documents préalablement filtrés ainsi que d'extensions filtrées subséquemment. Par ailleurs, nous mettons en œuvre un filtre basé sur la fréquence des documents au sein des extensions produites, dans le but de les épurer et de réduire la redondance des termes au travers de l'ensemble des documents. Pour conclure, l'expansion intègre un filtre global fondé sur la *SVD*, en privilégiant des termes proches des axes, afin de distinguer les termes caractéristiques du nouveau document par rapport à l'ensemble du corpus.

Nous comparons `LoGE` avec :

— La **Base**, qui se réfère exclusivement à la récupération effectuée par le biais de la fonction de score BM25 sur des documents ayant subi un pré-traitement à l'aide de filtres de base.

— Le modèle **Pegasus4IR**, pour lequel nous avons adapté l'approche `Pegasus` [Zha+20a] au domaine de la recherche d'informations. Dans ce cadre, nous procédons à la réécriture de textes en utilisant le modèle Pegasus pré-entraîné ayant dépassé les performances de l'état de l'art sur différents corpus pour la synthèse de documents.

— **UDEG** [Jeo+21] représente le modèle de référence dans le domaine de la génération abstraite d'extensions de documents destinées à la recherche ad hoc. Il constitue le modèle de l'état de l'art le plus proche, adoptant une démarche analogue en matière d'extension de documents de manière non supervisée.

Le tableau 3.3 illustre les performances des différentes approches sur les ensembles de données

TABLE 3.3 – Comparaison des performances de `LoGE` par rapport aux méthodes proches de l'état de l'art sur les corpus EU2UK et NFCorpus

	Modèle	MAP @10	F1-score @10	NDCG @10	Prec. @10	Rappel @10	Taille des extensions	Chevauchement de vocabulaire
EU2UK	Base	0,1607	0,1275	0,5609	0,0706	0,6814	4 473,64	0,2450
	Pegasus4IR	0,1719	**0,1538**	0,6058	**0,0853**	**0,8186**	7,20	0,0126
	UDEG	0,1618	0,1275	0,5625	0,0706	0,6814	4 491,12	0,2458
	LOGE	**0,2139**	**0,1538**	**0,7216**	**0,0853**	**0,8186**	959,32	0,0515
NFCorpus	Base	**0,1373**	0,1417	0,4534	0,0866	0,5848	2 121,04	0,2542
	Pegasus4IR	0,0812	0,0960	0,2766	0,0603	0,3741	16,68	0,0142
	UDEG	0,1368	0,1412	0,4529	0,0863	0,5838	2 161,26	0,2549
	LOGE	0,1370	**0,1442**	**0,4559**	**0,0884**	**0,5994**	795,06	0,0378

EU2UK et NFCorpus. Par rapport au tableau 3.2, l'introduction d'un filtre global se traduit par une amélioration significative des résultats. Dans le cas d'`EU2UK`, Pegasus4IR et `LoGE` affichent des performances comparables. Toutefois, notre méthode se distingue par la présentation de résultats mieux ordonnés, en raison de l'apport d'un plus grand nombre de termes dotés de nuances plus fines lors de l'évaluation des scores, particulièrement pour les documents volumineux. Sur `NFCorpus`, `LoGE` se révèle supérieur aux autres modèles en réduisant l'écart lexical. Pegasus4IR, limité par la brièveté des résumés générés qui ne contiennent pas un nombre suffisant de termes, ne parvient pas à s'adapter, tandis que UDEG est entravé par un excès de bruit.

En comparaison avec les modèles de référence non supervisés il est constaté que `LoGE` surpasse en termes de performances, y compris dans l'ordonnancement des résultats. Cette supériorité s'explique, d'une part, par une adaptabilité accrue aux documents de grande taille et, d'autre part, par l'intégration efficace de modèles spécifiques au domaine et prêts à l'emploi.

3.6 Conclusion

Dans ce chapitre, j'ai introduit `LoGE`, un moteur de recherche modulaire conçu pour intégrer de manière efficiente les contextes locaux et globaux des documents des corpus. Cette intégration s'effectue par le biais d'une génération de termes basée sur le fenêtrage et d'un filtre global fondé sur l'analyse en composantes singulières (SVD). Cette synergie offre un équilibre avantageux pour pallier les lacunes de vocabulaire dans des corpus textuels spécifiques et étendus. Elle contribue également à la réduction du bruit en synthétisant le contenu des documents à l'instar de la technique *SummVD* que j'ai présentée dans le chapitre 2 qui permet ainsi de résumer les documents. Dans le cas présent, il a permit de filtrer

les extensions de documents en se focalisant sur les synonymes les plus représentatifs.

Par ailleurs, notre méthode se caractérise par sa flexibilité et son autonomie, permettant l'intégration de modèles textuels préexistants (exemple : BERT) lesquels peuvent être adaptés à des domaines généraux ou spécifiques. Notre étude expérimentale a montré que les modèles spécifiques à un domaine exercent une influence plus prononcée lorsqu'il est nécessaire de générer des synonymes pertinents à ce domaine, tandis que les modèles généraux sont efficaces pour la paraphrase de documents.

Les travaux futurs devraient explorer l'efficacité temporelle de notre architecture par rapport à des approches similaires ainsi que les articles utilisant l'expansion des requêtes, car notre architecture de RI aide à faire évoluer le système. En outre, l'aspect modulaire de notre système pourrait être exploité en apprenant comment affiner le système. Nous pensons que notre approche constitue une base pour la composition de divers moteurs de recherche ad hoc spécialisés.

Troisième partie

Conclusion & perspectives

Chapitre 4

Conclusions et perspectives

4.1 Conclusion

Cette thèse s'est intéressée à la problématique émergente liée à l'augmentation exponentielle du volume des données textuelles et la nécessité indispensable pour les systèmes actuels de Recherche d'Information (RI) d'adopter des stratégies performantes et efficaces pour la sélection adéquate de l'ensemble documentaire en réponse à une requête spécifique. Un défi majeur identifié dans ce contexte est la divergence terminologique entre les termes employés dans les requêtes et ceux figurant dans les documents pertinents, une disparité sémantique particulièrement accentuée dans les documents volumineux issus de domaines spécialisés. De plus, il est impératif de développer des méthodes à faible coût, tant pour l'utilisateur que pour le temps d'entraînement nécessaire, afin de permettre une application industrielle viable des solutions proposées.

J'ai pu montrer que cette problématique était grandement accentuée lorsque nous nous intéressions à des corpus documentaires dédiés (juridiques ou médicaux) car le vocabulaire spécifique correspondant s'agrandit et augmente le bruit généré. Pour finir, je me suis intéressé au traitement de documents de grande taille présents dans ce contexte qui génèrent un bruit très important et ne facilitent pas la recherche d'information.

Dans cette optique, la thèse a introduit un système de RI non supervisé innovant, `LoGE`, s'appuyant sur des techniques avancées d'Apprentissage Profond (AP) telles que BERT et certaines de ses versions spécialisées au contexte considéré. Ce système, utilisant des modèles de langage masqués par AP, permet d'extrapoler des termes associés, enrichissant ainsi la représentation textuelle des

documents. Les modèles d'AP, formés sur de vastes corpus textuels, intègrent des connaissances générales ou spécifiques à un domaine, améliorant ainsi la représentation documentaire. Une projection des enrichissements, appelées extensions de documents, avec SVD permet de réduire le bruit généré et d'améliorer la qualité des résultats obtenus.

Le système `LoGE` a été rigoureusement évalué dans le domaine de la recherche ad-hoc sur des documents de grande envergure. Les résultats ont attesté de la capacité de généralisation du modèle et de son efficacité notable dans la production d'une représentation concise des mots et un enrichissement précis des documents, sans altérer le contexte global du document. Néanmoins, l'étude de l'état de l'art et les analyses des résultats initiaux autour de l'enrichissement de document a mis en évidence des enjeux significatifs en termes de rapport signal-bruit entre les requêtes et les documents pertinents. Il est donc apparu nécessaire de développer une méthode capable de minimiser le bruit dans les extensions générées tout en conservant leur pertinence.

L'approche *SummVD*, basée sur la décomposition en valeurs singulières, a été conçue pour réduire la dimensionnalité des plongements de mots et limiter l'étendue du vocabulaire, diminuant ainsi le bruit et augmentant l'efficacité du processus de synthèse textuelle. Sa faible consommation de ressources et sa rapidité d'exécution la rendent adaptée aussi bien à des documents uniques étendus qu'à d'importants corpus multidocumentaires, avec une application efficace en temps réel.

Les tests réalisés sur divers corpus ont confirmé la supériorité de *SummVD* par rapport aux méthodes extractives récentes, offrant une meilleure compréhension des problématiques liées à la dimension des documents et facilitant ainsi les méthodes de recherche d'information.

En associant les avantages de `LoGE` et son architecture modulaire à une version adaptée de *SummVD* pour la réduction du bruit dans les extensions générées, cette recherche a démontré des qualités de résultats remarquables, tout en maintenant un coût réduit en performance.

En conclusion, la présente thèse représente une contribution significative dans le champ de la RI, en proposant une stratégie innovante qui aborde efficacement les enjeux liés à la gestion des volumes en expansion de données textuelles dans le cadre spécifique de la recherche ad-hoc pour des documents longs. Cette méthode, tout en étant à la pointe de l'innovation, garde en ligne de mire les applications industrielles, répondant ainsi aux besoins des moteurs de recherche contemporains.

4.2 Perspectives

La méthode de résumé automatique proposée dans le chapitre 2 permet d'obtenir des performances sans adaptation spécifique des *plongements de mots* aux différents domaines. Les expérimentations effectuées sans connaissance a priori du domaine considéré ont produit des résultats pertinents malgré son absence de spécialisation. Il serait intéressant d'étudier l'impact de modèles pré-entraînés avec un vocabulaire spécifique, comme le secteur médical, scientifique ou encore les médias sociaux. Ainsi, lors de l'étape de vectorisation du document nettoyé, le modèle pré-entraîné produit un espace vectoriel spécialisé au contexte donné et serait à même de produire des partitionnements différents d'un modèle générique. En conséquence, la réduction SVD produirait des mots significatifs plus spécifiques et aurait une incidence sur les scores lors de la sélection des phrases représentant le document. Le résumé résultant serait ainsi plus pertinent et spécifique.

Par ailleurs, la faible complexité et la flexibilité de l'approche *SummVD* vis-à-vis de la nature et de la taille des documents suggèrent de vastes possibilités d'exploration sur de grands ensembles de données multi documents, à l'image de ceux de *Google, TripAdvisor* ou encore *Amazon*. En effet, il serait intéressant de résumer de larges corpus de documents en exploitant différentes dimensions. Prenons l'exemple de données touristiques comme *TripAdvisor* : l'aspect temporel pourrait être exploité en analysant l'évolution du résumé au cours du temps sur une localité, la vision globale d'une destination en agrégeant les commentaires de touristes à propos de plusieurs sous-ensembles d'une destination qui favoriseront la représentativité de celle-ci en réduisant l'impact d'un lieu précis de la zone considérée.

Les orientations futures de la recherche devraient se concentrer sur l'évaluation comparative de l'efficacité temporelle de l'architecture `LoGE` par rapport à d'autres architectures similaires, ainsi que par rapport aux systèmes qui implémentent l'expansion des requêtes. Cette exploration est cruciale pour établir la supériorité ou l'efficacité relative de notre architecture dans le contexte de besoin d'efficacité. Les essais de performance requis pour évaluer notre architecture `LoGE` , de nature modulaire, peuvent être réalisés sans souci majeur. En effet, l'intégration des modèles de pointe dans le cadre de l'interrogation de la base de données ne présente pas de complexité majeure au sein de notre structure.

Bien que nécessaires pour parfaire l'approche proposée, j'ai préféré focaliser mes expérimentations sur la qualité des résultats obtenus lors de la génération d'extensions de documents plutôt que sur

les temps de réponse. Une architecture se basant sur *Elasticsearch* avec des listes inverses optimisées constitue un socle architectural inégalable en matière de recherche d'information. La facilité de mise en œuvre et sa résilience ont été un point de décision important dans la conception de l'architecture.

Pour valider notre hypothèse d'efficacité, il s'agira de constituer un ensemble de données de test standardisé, qui sera employé pour évaluer chacun des modèles concernés. Cela permettra de mesurer et de comparer le temps de réponse des modèles actuels, ceux utilisant des expansions de requêtes, avec notre système lors de l'exécution de requêtes dans la base de données. Il est nécessaire de rappeler que les techniques d'expansion de requêtes nécessitent un traitement de chaque requête, tandis que `LoGE` effectue cette étape en pré-traitement. Ainsi, cette approche méthodique garantira une évaluation précise et fiable de l'efficacité temporelle de notre architecture en comparaison avec les solutions existantes dans le domaine de la recherche d'information.

En outre, l'aspect modulaire de notre système pourrait être exploité en apprenant comment affiner le système. Nous pensons que notre approche constitue une base pour la composition de divers moteurs de recherche ad hoc spécialisés.

Notamment, l'optimisation du processus de fenêtrage, présenté dans la section 3.3.3.1, dans la génération d'extensions de texte mérite une attention particulière. Actuellement, cette procédure est principalement restreinte aux dimensions des couches d'entrée des divers modèles de langage à base de masque.

Cette méthode de découpage systématique du texte, qui ne tient pas compte de la structure globale du document, peut conduire à une juxtaposition inadéquate d'informations issues de différentes sections du document au sein du modèle. En effet, dans le cas de documents de grande longueur, où différentes sections peuvent aborder des sujets distincts, cette approche peut s'avérer problématique. En particulier, lors de la génération d'extensions pour des mots situés aux frontières de ces sections, les prédictions générées peuvent être affectées, résultant en la suggestion de termes non pertinents ou hors sujet. Par conséquent, il devient impératif d'élaborer des stratégies de fenêtrage plus sophistiquées qui prennent en compte la structure et le contenu global du document (phrases, sections, chapitres, etc.) pour améliorer la précision et la pertinence des extensions générées dans ces modèles de langage.

De manière similaire à la méthode *SummVD*, il est possible de générer des plongements vectoriels pour les phrases ou les termes individuels en utilisant un modèle approprié, créant ainsi un espace

vectoriel pour chaque document. Cette approche facilite la visualisation de la distance sémantique entre les différents segments analysés. Un partitionnement des données peut être effectué de manière non supervisée afin d'identifier les thèmes majeurs présents dans le document et de segmenter ce dernier en fonction de l'appartenance des phrases consécutives à chacun des groupes thématiques non explicitement nommés. Cette stratégie est particulièrement avantageuse pour fournir des segments cohérents aux modèles de langage utilisés dans `LoGE`, enrichis d'informations uniformes et précises, ce qui contribue à l'amélioration de la pertinence des prédictions générées par ces modèles. Une telle méthode représente donc un outil efficace pour structurer et clarifier les données textuelles en vue d'une génération plus affinée des extensions.

Bibliographie

[Als+19] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann et Matthew McDermott. « Publicly Available Clinical BERT Embeddings ». In : *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA : Association for Computational Linguistics, juin 2019, p. 72-78. doi : 10.18653/v1/W19-1909.

[Ank+99] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel et Jörg Sander. « OPTICS : Ordering Points to Identify the Clustering Structure ». In : *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*. T. 28. 2. New York, NY, USA : Association for Computing Machinery, juin 1999, 49–60. doi : 10.1145/304181.304187.

[ART20] Oussama Ayoub, Christophe Rodrigues et Nicolas Travers. « Adaptive Search Engine for Heterogeneous Documents ». In : *Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA'20)*. 4685. Paris, France, oct. 2020. url : https://easychair.org/publications/preprint_open/wFSS.

[ART22] Oussama Ayoub, Christophe Rodrigues et Nicolas Travers. « Un générateur d'extension de documents non supervisé pour moteurs de recherche ». In : *GdR Traitement Automatique de la Langue*. CNRS - IRISA. Rennes, 1er oct. 2022. url : https://gdr-tal-rennes.sciencesconf.org/resource/page/id/2.

[ART23] Oussama Ayoub, Christophe Rodrigues et Nicolas Travers. « LoGE : an unsupervised local-global document extension generation in information retrieval for long documents ». In : *International Journal of Web Information Systems* 19.5/6 (nov. 2023), p. 244-262. doi : 10.1108/IJWIS-07-2023-0109.

[ATM14] Emhimed Alatrish, Dušan Tošić et Nikola Milenkovic. « Building ontologies for different natural languages ». In : *Computer Science and Information Systems* 11.2 (juin 2014), p. 623-644. doi : 10.2298/CSIS130429023A.

[Ayo+23] Oussama Ayoub, Ludovic Li, Christophe Rodrigues et Nicolas Travers. « LoGE : Expansion Locale-Globale de document non supervisée avec un moteur de recherche Extensible ». In : *TextMine Workshop @ EGC'23*. Lyon, France, 2023, p. 41-44. url : https://textmine.sciencesconf.org/data/pages/TextMine23.pdf.

[Bar+15] Federico Barrios, Federico López, Luis Argerich et Rosa Wachenchauzer. « Variations of the Similarity Function of TextRank for Automated Summarization ». In : *The 16th Argentine Symposium on Artificial Intelligence*. Rosario, 2015, p. 65-72. url : https://44jaiio.sadio.org.ar/sites/default/files/asai65-72.pdf.

[BME99] Regina Barzilay, Kathleen R. McKeown et Michael Elhadad. « Information Fusion in the Context of Multi-Document Summarization ». In : *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL '99. College Park, Maryland : Association for Computational Linguistics, 1999, 550–557. isbn : 1558606093. doi : 10.3115/1034678.1034760.

[Bot+16] Vera Boteva, Demian Gholipour, Artem Sokolov et Stefan Riezler. « A Full-Text Learning to Rank Dataset for Medical Information Retrieval ». In : *European Conference on Information Retrieval (ECIR'16)*. Padua, Italy : Springer International Publishing, 2016, p. 716-722. isbn : 978-3-319-30671-1. doi : https://doi.org/10.1007/978-3-319-30671-1_58.

[BR17] Aurélien Bossard et Christophe Rodrigues. « An Evolutionary Algorithm for Automatic Summarization ». In : *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Varna, Bulgaria : INCOMA Ltd., sept. 2017, p. 111-120. doi : 10.26615/978-954-452-049-6_017.

[BZ05] Bodo Billerbeck et Justin Zobel. « Document expansion versus query expansion for ad-hoc retrieval ». In : *ADCS 2005 - Proceedings of the Tenth Australasian Document Computing Symposium* (déc. 2005). url : https://core.ac.uk/outputs/15628607.

[CD22] Shubham Chatterjee et Laura Dietz. « BERT-ER : Query-Specific BERT Entity Representations for Entity Ranking ». In : *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. Madrid, Spain, 2022, 1466–1477. isbn : 9781450387323. doi : 10.1145/3477495.3531944.

[CG98] Jaime Carbonell et Jade Goldstein. « The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries ». In : *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '98. Melbourne, Australia : Association for Computing Machinery, 1998, 335–336. isbn : 1581130155. doi : 10.1145/290941.291025.

[Cha+20] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras et Ion Androutsopoulos. « LEGAL-BERT : The Muppets straight out of Law School ». In : *Conference on Empirical Methods in Natural Language Processing (EMNLP'20), online*. 2020, p. 2898-2904. doi : 10.18653/v1/2020.findings-emnlp.261.

[Cha+21] Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalou et Prodromos Malakasiotis. « Regulatory Compliance through Doc2Doc Information Retrieval : A case study in EU/UK legislation where text similarity has limitations ». In : *Conference of the European Chapter of the Association for Computational Linguistics (EACL'21), online*. 2021. doi : 10.18653/v1/2021.eacl-main.305.

[Chu17] Kenneth Ward Church. « Word2Vec ». In : *Natural Language Engineering* 23.1 (2017). Sous la dir. de Cambridge University Press, 155–162. doi : 10.1017/S1351324916000334.

[Coh+18] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang et Nazli Goharian. « A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents ». In : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana : Association for Computational Linguistics, juin 2018, p. 615-621. doi : 10.18653/v1/N18-2097.

[Dee+90] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer et Richard Harshman. « Indexing by latent semantic analysis ». In : *Journal of the American Society for Information Science* 41.6 (1990), p. 391-407. doi : (SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[Dev+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee et Kristina Toutanova. « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding ». In : *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota : Association for Computational Linguistics, juin 2019, p. 4171-4186. doi : 10.18653/v1/N19-1423.

[Edm69] Harold P. Edmundson. « New Methods in Automatic Extracting ». In : *Journal of ACM* 16.2 (avr. 1969), 264–285. issn : 0004-5411. doi : 10.1145/321510.321519.

[Est+96] Martin Ester, Hans-Peter Kriegel, Jörg Sander et Xiaowei Xu. « A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise ». In : *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon : AAAI Press, 1996, 226–231. doi : 10.5555/3001460.3001507.

[Fab+19] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li et Dragomir Radev. « Multi-News : A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model ». In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy : Association for Computational Linguistics, juill. 2019, p. 1074-1084. doi : 10.18653/v1/P19-1102.

[Far18] Mamdouh Farouk. « Sentence Semantic Similarity based on Word Embedding and WordNet ». In : *2018 13th International Conference on Computer Engineering and Systems (ICCES)*. 2018, p. 33-37. doi : 10.1109/ICCES.2018.8639211.

[For65] E. W. Forgy. « Cluster Analysis of Multivariate Data : Efficiency versus Interpretability of Classification ». In : *Biometrics* 21.3 (1965), p. 768-769. url : https://cir.nii.ac.jp/crid/1571980074621944832.

[GF09] Dan Gillick et Benoit Favre. « A Scalable Global Model for Summarization ». In : *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*. Boulder, Colorado : Association for Computational Linguistics, juin 2009, p. 10-18. url : https://aclanthology.org/W09-1802.

[GL01] Yihong Gong et Xin Liu. « Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis ». In : *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '01. New Orleans, Louisiana, USA : Association for Computing Machinery, 2001, 19–25. isbn : 1581133316. doi : 10.1145/383952.383955.

[Gon+18] Hongyu Gong, Tarek Sakakini, Suma Bhat et JinJun Xiong. « Document Similarity for Texts of Varying Lengths via Hidden Topics ». In : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Melbourne, Australia : Association for Computational Linguistics, juill. 2018, p. 2341-2351. doi : 10.18653/v1/P18-1218.

[Gra15] Yvette Graham. « Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE ». In : *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal : Association for Computational Linguistics, sept. 2015, p. 128-137. DOI : 10.18653/v1/D15-1013.

[GVL96] Gene H. Golub et Charles F. Van Loan. *Matrix Computations*. Third. The Johns Hopkins University Press, 1996.

[Her+15] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman et Phil Blunsom. « Teaching Machines to Read and Comprehend ». In : *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada : MIT Press, 2015, 1693–1701. URL : https://dl.acm.org/doi/10.5555/2969239.2969428.

[Jal+04] Nasreen Jaleel, James Allan, W. Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark Smucker et Courtney Wade. « UMass at TREC 2004 : Novelty and hard ». In : *Text Information Retrieval Conference (TREC'04)*. Jan. 2004. URL : https://scholarworks.umass.edu/cs_faculty_pubs/189/.

[Jeo+21] Soyeong Jeong, Jinheon Baek, ChaeHun Park et Jong Park. « Unsupervised Document Expansion for Information Retrieval with Stochastic Text Generation ». In : *Proceedings of the Second Workshop on Scholarly Document Processing*. online : Association for Computational Linguistics, juin 2021, p. 7-17. DOI : 10.18653/v1/2021.sdp-1.2.

[Kee+19] Robert Keeling, Rishi Chhatwal, Nathaniel Huber-Fliflet, Jianping Zhang, Fusheng Wei, Haozhen Zhao, Ye Shi et Han Qin. « Empirical Comparisons of CNN with Other Learning Algorithms for Text Classification in Legal Document Review ». In : *2019 IEEE International Conference on Big Data (Big Data)*. Los Angeles, CA, USA, 2019, p. 2038-2042. DOI : 10.1109/BigData47090.2019.9006248.

[LC12] Qiang Lu et Jack G. Conrad. « Bringing Order to Legal Documents - An Issue-based Recommendation System Via Cluster Association ». In : *International Conference on Knowledge Engineering and Ontology Development*. Barcelona, Spain, 2012, p. 76-88. URL : http://www.conradweb.org/~jackg/pubs/KEOD12_Lu_Conrad_STP.pdf.

[Li+23] Canjia Li, Andrew Yates, Sean MacAvaney, Ben He et Yingfei Sun. « PARADE : Passage Representation Aggregation ForDocument Reranking ». In : *ACM Trans. Inf. Syst.* 42.2 (sept. 2023). ISSN : 1046-8188. DOI : 10.1145/3600088.

[Lin04] Chin-Yew Lin. « ROUGE : A Package for Automatic Evaluation of Summaries ». In : *Text Summarization Branches Out*. Barcelona, Spain : Association for Computational Linguistics, juill. 2004, p. 74-81. URL : https://aclanthology.org/W04-1013.

[Liu+21] Ye Liu, Jianguo Zhang, Yao Wan, Congying Xia, Lifang He et Philip Yu. « HETFORMER : Heterogeneous Transformer with Sparse Attention for Long-Text Extractive Summarization ». In : *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online et Punta Cana, Dominican Republic : Association for Computational Linguistics, nov. 2021, p. 146-154. DOI : 10.18653/v1/2021.emnlp-main.13.

[Luh58] Hans Peter Luhn. « The Automatic Creation of Literature Abstracts ». In : *IBM Journal of Research and Development* 2.2 (avr. 1958), 159–165. ISSN : 0018-8646. DOI : 10.1147/rd.22.0159.

[Ma+22] Xueguang Ma, Ronak Pradeep, Rodrigo Nogueira et Jimmy Lin. « Document Expansion Baselines and Learned Sparse Lexical Representations for MS MARCO V1 and V2 ». In : *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. Madrid, Spain, 2022, 3187–3197. isbn : 9781450387323. doi : 10.1145/3477495.3531749.

[Mik+13] Tomás Mikolov, Kai Chen, Greg Corrado et Jeffrey Dean. « Efficient Estimation of Word Representations in Vector Space ». In : *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. 2013. arXiv : abs/1301.3781.

[MRS08] Christopher D. Manning, Prabhakar Raghavan et Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. url : http://nlp.stanford.edu/IR-book/information-retrieval-book.html.

[MT04] Rada Mihalcea et Paul Tarau. « TextRank : Bringing Order into Text ». In : *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain : Association for Computational Linguistics, juill. 2004, p. 404-411. url : https://aclanthology.org/W04-3252.

[Nal+16] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre et Bing Xiang. « Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond ». In : *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany : Association for Computational Linguistics, août 2016, p. 280-290. doi : 10.18653/v1/K16-1028.

[Nas+21] Shahrzad Naseri, Jeffrey Dalton, Andrew Yates et James Allan. « CEQE : Contextualized Embeddings for Query Expansion ». In : *Advances in Information Retrieval*. Cham : Springer International Publishing, 2021, p. 467-482. isbn : 978-3-030-72113-8. doi : 10.1007/978-3-030-72113-8_31.

[NCL18] Shashi Narayan, Shay B. Cohen et Mirella Lapata. « Don't Give Me the Details, Just the Summary ! Topic-Aware Convolutional Neural Networks for Extreme Summarization ». In : *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium : Association for Computational Linguistics, oct. 2018, p. 1797-1807. doi : 10.18653/v1/D18-1206.

[Nog+19] Rodrigo Nogueira, Wei Yang, Jimmy Lin et Kyunghyun Cho. « Document Expansion by Query Prediction ». In : *CoRR* abs/1904.08375 (2019). arXiv : 1904.08375.

[Nog+20] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep et Jimmy Lin. « Document Ranking with a Pretrained Sequence-to-Sequence Model ». In : *Findings of the Association for Computational Linguistics : EMNLP 2020*. Online : Association for Computational Linguistics, nov. 2020, p. 708-718. doi : 10.18653/v1/2020.findings-emnlp.63.

[NW17] Adeline Nazarenko et Adam Wyner. « Legal NLP Introduction ». In : *Traitement Automatique des Langues* 58".2 (2017), p. 7-19. url : https://aclanthology.org/2017.tal-2.1.

[OCM17] Jessica Ouyang, Serina Chang et Kathy McKeown. « Crowd-Sourced Iterative Annotation for Narrative Summarization Corpora ». In : *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*. Valencia, Spain : Association for Computational Linguistics, avr. 2017, p. 46-51. URL : https://aclanthology.org/E17-2008.

[Ped+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot et Édouard Duchesnay. « Scikit-learn : Machine Learning in Python ». In : *Journal of Machine Learning Research* 12.85 (2011), p. 2825-2830. URL : http://jmlr.org/papers/v12/pedregosa11a.html.

[Pfe+18] Jonas Pfeiffer, Samuel Broscheit, Rainer Gemulla et Mathias Göschl. « A Neural Autoencoder Approach for Document Ranking and Query Refinement in Pharmacogenomic Information Retrieval ». In : *Proceedings of the BioNLP 2018 workshop*. Melbourne, Australia : Association for Computational Linguistics, juill. 2018, p. 87-97. DOI : 10.18653/v1/W18-2310.

[PH21] Vishakh Padmakumar et He He. « Unsupervised Extractive Summarization using Pointwise Mutual Information ». In : *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*. Online : Association for Computational Linguistics, avr. 2021, p. 2505-2512. DOI : 10.18653/v1/2021.eacl-main.213.

[PSM14] Jeffrey Pennington, Richard Socher et Christopher Manning. « GloVe : Global Vectors for Word Representation ». In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar : Association for Computational Linguistics, août 2014, p. 1532-1543. DOI : 10.3115/v1/D14-1162.

[RJ71] Joseph John Rocchio Jr. « Relevance Feedback in Information Retrieval ». In : *The SMART retrieval system : experiments in automatic document processing* (1971). URL : https://sigir.org/files/museum/pub-08/XXIII-1.pdf.

[RJ76] Stephen Robertson et K. Sparck Jones. « Relevance weighting of search terms ». In : *Journal of the American Society for Information Science* 27 (jan. 1976), p. 129-146. URL : https://www.microsoft.com/en-us/research/publication/relevance-weighting-of-search-terms/.

[RJB00] Dragomir R. Radev, Hongyan Jing et Malgorzata Budzikowska. « Centroid-based summarization of multiple documents : sentence extraction, utility-based evaluation, and user studies ». In : *NAACL-ANLP 2000 Workshop : Automatic Summarization*. Seattle Washington, 2000. URL : https://aclanthology.org/W00-0403.

[ŘS10] Radim Řehůřek et Petr Sojka. « Software Framework for Topic Modelling with Large Corpora ». English. In : *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta : ELRA, mai 2010, p. 45-50.

[Ryg+17] Jan Rygl, Jan Pomikálek, Radim Řehůřek, Michal Růžička, Vít Novotný et Petr Sojka. « Semantic Vector Encoding and Similarity Search Using Fulltext Search Engines ». In : *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Vancouver, Canada : Association for Computational Linguistics, août 2017, p. 81-90. doi : 10.18653/v1/W17-2611.

[Sha+20] Xuan Shan, Chuanjie Liu, Yiqian Xia, Qi Chen, Yusi Zhang, Angen Luo et Yuxiang Luo. « BISON : BM25-weighted Self-Attention Framework for Multi-Fields Document Search ». In : *CoRR* abs/2007.05186 (2020). arXiv : abs/2007.05186 [cs.IR].

[She+22] Gabriel Shenouda, Aurélien Bossard, Oussama Ayoub et Christophe Rodrigues. « SummVD : An efficient approach for unsupervised topic-based text summarization ». In : *International Joint Conference on Natural Language Processing (IJCNLP'22), online*. Nov. 2022, p. 501-511. url : https://aclanthology.org/2022.aacl-main.38.

[SJ05] Josef Steinberger et Karel Ježek. « Text Summarization and Singular Value Decomposition ». In : *Advances in Information Systems*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2005, p. 245-254. isbn : 978-3-540-30198-1. url : https://link.springer.com/chapter/10.1007/978-3-540-30198-1_25.

[SLM17] Abigail See, Peter J. Liu et Christopher D. Manning. « Get To The Point : Summarization with Pointer-Generator Networks ». In : *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. Vancouver, Canada : Association for Computational Linguistics, juill. 2017, p. 1073-1083. doi : 10.18653/v1/P17-1099.

[SWY75] Gerard M. Salton, Andrew Wong et Chungshu Yang. « A Vector Space Model for Automatic Indexing ». In : *Commun. ACM* 18.11 (nov. 1975), 613–620. issn : 0001-0782. doi : 10.1145/361219.361220.

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser et Illia Polosukhin. « Attention is All You Need ». In : *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA : Curran Associates Inc., 2017, 6000–6010. isbn : 9781510860964. url : https://dl.acm.org/doi/10.5555/3295222.3295349.

[WLA22] Yumeng Wang, Lijun Lyu et Avishek Anand. « BERT Rankers Are Brittle : A Study Using Adversarial Document Perturbations ». In : *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*. ICTIR '22. Madrid, Spain, 2022, 115–120. isbn : 9781450394123. doi : 10.1145/3539813.3545122.

[Wu+21a] Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Ziqiang Cao, Sujian Li, Hua Wu et Haifeng Wang. « BASS : Boosting Abstractive Summarization with Unified Semantic Graph ». In : *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*. Online : Association for Computational Linguistics, août 2021, p. 6052-6067. doi : 10.18653/v1/2021.acl-long.472.

[Wu+21b] Ye Wu, Hing-Fung Ting, Tak-Wah Lam et Ruibang Luo. « BioNumQA-BERT : Answering Biomedical Questions Using Numerical Facts with a Deep Language Representation Model ». In : *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. BCB '21. Gainesville, Florida, 2021. ISBN : 9781450384506. DOI : 10.1145/3459930.3469557.

[Xio+17] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu et Russell Power. « End-to-End Neural Ad-hoc Ranking with Kernel Pooling ». In : *Conference on Research and Development in Information Retrieval (SIGIR'17)*. Tokyon, Japan, 2017, p. 55-64. DOI : 10.1145/3077136.3080809.

[Zer+22] George Zerveas, Navid Rekabsaz, Daniel Cohen et Carsten Eickhoff. « Mitigating Bias in Search Results Through Contextual Document Reranking and Neutrality Regularization ». In : *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. Madrid, Spain, 2022, 2532–2538. ISBN : 9781450387323. DOI : 10.1145/3477495.3531891.

[Zha+20a] Jingqing Zhang, Yao Zhao, Mohammad Saleh et Peter J. Liu. « PEGASUS : Pre-Training with Extracted Gap-Sentences for Abstractive Summarization ». In : *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. Online : JMLR.org, juill. 2020, 11328–11339. URL : https://dl.acm.org/doi/pdf/10.5555/3524938.3525989.

[Zha+20b] Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang et Gholamreza Haffari. « SummPip : Unsupervised Multi-Document Summarization with Sentence Graph Compression ». In : *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '20. Virtual Event, China : Association for Computing Machinery, 2020, 1949–1952. ISBN : 9781450380164. DOI : 10.1145/3397271.3401327.

[Zhe+20] Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun et Andrew Yates. « BERT-QE : Contextualized Query Expansion for Document Re-ranking ». In : *Findings of the Association for Computational Linguistics : EMNLP 2020*. Online : Association for Computational Linguistics, nov. 2020, p. 4718-4728. DOI : 10.18653/v1/2020.findings-emnlp.424. URL : https://aclanthology.org/2020.findings-emnlp.424.

[Zho+20] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu et Xuanjing Huang. « Extractive Summarization as Text Matching ». In : *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online : Association for Computational Linguistics, juill. 2020, p. 6197-6208. DOI : 10.18653/v1/2020.acl-main.552.

[Zhu+21] Honglei Zhuang, Zhen Qin, Shuguang Han, Xuanhui Wang, Michael Bendersky et Marc Najork. « Ensemble Distillation for BERT-Based Ranking Models ». In : *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. ICTIR '21. Virtual Event, Canada, 2021, 131–136. ISBN : 9781450386111. DOI : 10.1145/3471158.3472238.

[Zhu+22] Tiange Zhu, Raphaël Fournier-S'niehotta, Philippe Rigaux et Nicolas Travers. « A Framework for Content-Based Search in Large Music Collections ». In : *Big Data and Cognitive Computing* 6.1 (2022), p. 23. DOI : 10.3390/bdcc6010023.

[Zhu+23] Wenhao Zhu, Xiaoyu Zhang, Liang Ye et Qiuhong Zhai. « Query Context Expansion for Open-Domain Question Answering ». In : *ACM Transactions on Asian and Low-Resource Language Information Processing* 22.8 (août 2023), p. 1-21. ISSN : 2375-4699. DOI : 10.1145/3603498. URL : https://doi.org/10.1145/3603498.