كلية العلوم و التقنيات فاس
ⵜⴰⵖⵙⵓⵍⵜ ⵏ ⵜⵇⵇⵓⵔⵉⵙⵏ ⴷ ⵜⵓⵙⵙⵙⵜⵙⵏ
Faculté des Sciences et Techniques de Fès

جامعة سيدي محمد بن عبد الله
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵓⵙⴰⵙ ⵎⵃⵎⴷ ⴱⵏ ⵄⴱⴷⵍⵍⴰⵀ
Université Sidi Mohamed Ben Abdellah

Centre d'Etudes Doctorales : Sciences et Techniques de l'Ingénieur

N : 86 / 2019

THÈSE DE DOCTORAT

Présentée Par

Mohammed RAIS

Directeur de thèse
Prof. Abdelmonaime LACHKAR

Discipline : Sciences de l'Ingénieur

Spécialité : Informatique

Étude et Amélioration de la Représentation du Texte Biomédical Basée su
l'Intégration des Ontologies du Domaine et l'Enrichissement Sémantique

---

Soutenue le 09 / 11 / 2019, Devant le jury :

Nom Prénom	Titre	Etablissement	
Mohammed Ouçamah CHERKAOUI MALKI	PES	Faculté des Sciences Dhar El Mehraz, USMBA, Fès	Président
Majda AICHA	PH	Faculté des Sciences et Techniques de Fès, USMBA, Fès	Rapporteur
Zineb SERHIER	PES	Faculté de Médecine et de Pharmacie, UH II, Casablanca	Rapporteur
Abderrahim ELQADI	PES	Ecole Supérieure de Technologie de Salé, UM5, Rabat	Rapporteur
Noureddine CHENFOUR	PES	Faculté des Sciences Dhar El Mehraz, USMBA, Fès	Examinateur
Abdelmonaime LACHKAR	PES	Ecole Nationale des Sciences Appliquées, UAE, Tanger	Directeur de thèse

Laboratoire d'Accueil : Ingénierie Systèmes et Applications

Etablissement : Ecole Nationale des Sciences Appliquées de Fès

بسم الله الرحمن الرحيم

"وَلَقَدْ خَلَقْنَا الْإِنسَانَ مِن سُلَالَةٍ مِّن طِينٍ (12) ثُمَّ جَعَلْنَاهُ نُطْفَةً فِي قَرَارٍ مَّكِينٍ (13) ثُمَّ خَلَقْنَا النُّطْفَةَ عَلَقَةً فَخَلَقْنَا الْعَلَقَةَ مُضْغَةً فَخَلَقْنَا الْمُضْغَةَ عِظَامًا فَكَسَوْنَا الْعِظَامَ لَحْمًا ثُمَّ أَنشَأْنَاهُ خَلْقًا آخَرَ ۚ فَتَبَارَكَ اللَّهُ أَحْسَنُ الْخَالِقِينَ (14) ثُمَّ إِنَّكُم بَعْدَ ذَٰلِكَ لَمَيِّتُونَ (15) ثُمَّ إِنَّكُمْ يَوْمَ الْقِيَامَةِ تُبْعَثُونَ (16)"

صدق الله العظيم

(سورة المؤمنون، الآية 16 - 12)

إهداء

إلهي لا يطيب الليل إلا بشكرك ولا يطيب النهار إلى بطاعتك... ولا تطيب اللحظات إلا بذكرك... ولا تطيب الآخرة إلا بعفوك... ولا تطيب الجنة إلا برؤيتك

إلى من بلغ الرسالة وأدى الأمانة... ونصح الأمة... إلى نبي الرحمة ونور العالمين

سيدنا محمد صلى الله عليه وسلم

إلى ملاكي في الحياة... إلى معنى الحب وإلى معنى الحنان والتفاني... إلى التي حملتني وهنا بسمة الحياة إلى شمعة متقدة تنير ظلمة حياتي...
إلى من بوجودها أكتسب قوة ومحبة لا حدود لها...
إلى من كان حنانها بلسم جراحي

أمي الحبيبة

إلى من كلله الله بالهيبة والوقار... إلى من علمني العطاء بدون انتظار... إلى من أحمل أسمه بكل افتخار... أرجو من الله أن يمد في عمرك لترى ثماراً قد حان قطافها بعد طول انتظار وستبقى كلماتك نجوم أهتدي بها اليوم وفي الغد وإلى الأبد

والدي العزيز

إلى من بوجودهما أكتسب قوة ومحبة لا حدود لها...
إلى من عرفت معهما معنى الحياة

زوجتي وابنتي

إلى جدتي فاطمة الهاشمي وجدتي زينب الودغيري أطال الله في عمرها

إلى من حبهم يجري في عروقي ويلهج بذكراهم فؤادي... إلى أخواتي وأخواني سعدية، شيماء، هاجر، يوسف وفتيحة

إلى جميع أفراد عائلتي

إلى كل من علمني حرفا

أهدي هذا البحث المتواضع راجياً من المولى

عز وجل أن يجد القبول والنجاح

Remerciements

Je tiens tout d'abord à remercier Allah, le tout puissant, pour cette grâce d'être en vie et en bonne santé depuis ma naissance jusqu'à ce jour, et qui m'a donné la force, le courage et la patience d'accomplir ce Modeste travail.

Le travail de cette thèse est le produit de plusieurs années de recherche durant lesquelles de nombreuses personnes ont marqué leurs empruntes clairement dans les traces réalisées pendant cette période et que je tiens à les remercier.

Je tiens à remercier spécialement mon directeur de thèse, Monsieur **Abdelmonaime LACHKAR**, professeur à l'Ecole Nationale des Sciences Appliquées de Tanger (Ex-Professeur à ENSA-USMBA, Fès), qui a encadré et dirigé de près le présent travail de thèse. Je vous remercie pour le temps et la patience que vous m'avez accordée, pour vos conseils fructueux que vous m'avez prodigués, pour votre orientation ciblée, pour votre disponibilité, même à des heures peu adaptées tout au long de ces années. Pour tout ce que vous m'avez donné, je vous remercie très sincèrement.

Merci également aux membres de jury, pour avoir accepté de juger ce travail auquel ils ont accordé un intérêt particulier.

Je remercie toutes les personnes qui ont participé de manière directe ou indirecte à la concrétisation de ce travail. Qu'ils trouvent ici l'expression de ma reconnaissance.

Publications de l'Auteur

Publications

[1] **Mohammed RAIS**, Issam SAHMOUDI et Abdelmonaime LACHKAR, "Formal Concept Analysis for Biomedical Web Search Results Clustering ", soumis, Mai 2019, "***Journal of Computer Science****"*

[2] **Mohammed RAIS**, Mohammed BEKKALI et Abdelmonaime LACHKAR, "An Efficient Method for Biomedical Word Sense Disambiguation Based On New Web-Kernel Similarity", Accepté pour publication, Mai 2019, "***The International Journal of Healthcare Information Systems and Informatics (IJHISI)****"*

[3] **Mohammed RAIS** et Abdelmonaime LACHKAR, "An Empirical Study of Word Sense Disambiguation for Biomedical Information Retrieval System", 2018, Book: "***Bioinformatics and Biomedical Engineering Lecture Notes in bioinformatics (LNBI)***" – LNCS vol 10813. DOI: 10.1007/978-3-319-78723-7_27

[4] **Mohammed RAIS** et Abdelmonaime LACHKAR, "Evaluation of Disambiguation Strategies on Biomedical Text Categorization", 2016, Book: "***Bioinformatics and Biomedical Engineering Lecture Notes in bioinformatics (LNBI)***" – LNCS vol 9656. DOI: 10.1007/978-3-319-31744-1_68

Conférences

[1] **Mohammed RAIS** et Abdelmonaime LACHKAR, "An Empirical Study of Word Sense Disambiguation for Biomedical Information Retrieval System", The 6th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2018). 25 – 27 Avril 2018, Granada, Spain.

[2] **Mohammed RAIS** et Abdelmonaime LACHKAR, "Evaluation of Disambiguation Strategies on Biomedical Text Categorization", The 4th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2016). 20 – 22 Avril 2016, Granada, Spain.

[3] **Mohammed. RAIS** et Abdelmonaime LACHKAR, "Resolving Ambiguity in Biomedical Text Representation to Improve Text Categorization", Workshop du Printemps Lisa-WP2016". 21 Avril 2016, Fès, Maroc.

[4] **Mohammed RAIS** et Abdelmonaime LACHKAR, "Biomedical Word Sense Disambiguation Context-Based: Improvement of SenseRelate Method", IEEE 2nd International Conference on Information Technology for Organizations Development (IT4OD-2016). 30 Mars – 1 Avril 2016, Fès, Maroc.

[5] **Mohammed RAIS**, Abdelmonaime LACHKAR, Abdelhamid LACHKAR et S. EL ALAOUI OUATIK: "A Comparative Study of Biomedical Named Entity Recognition Methods Based Machine Learning Approach", IEEE Information and Multimedia Processing CIST 2014" 20-22 October 2014 – Tetouan, Maroc.

[6] Abdelhamid LACHKAR, **Mohammed RAIS**, S. EL ALAOUI OUATIK et Abdelmonaime LACHKAR: "Performance Evaluation of Markov Model Based Methods for Biomedical Named Entity Recognition", Journées Doctorales en Technologies de l'Information et de la Communication (JDTIC'14)" 19-20 Juin 2014 – Rabat, Maroc.

الملخص

في العقود الأخيرة، نمت كمية المعلومات النصية بشكل كبير في جميع المجالات، بما في ذلك المجال الطبي والحيوي. وقد أصبح الوصول إلى المعلومات في مجال الطب الحيوي أكثر صعوبة. لذلك من المهم للغاية التعامل مع هذه المشكلة، واقتراح الأساليب، وإعداد أدوات لمساعدة مستخدمي المعلومات في مجال الطب الحيوي للوصول إليها بسرعة وسهولة. من أجل تحسين أداء نتائج أي تطبيق لتحليل النصوص الطبية الحيوية، في إطار هذه الأطروحة، نقترح بشكل أساسي دراسة مشكلة تمثيل النص الطبي الحيوي ونتعامل مع تحديات فهرسة المستندات الطبية الحيوية. بالإضافة إلى ذلك، نقترح أيضًا دراسة مشكلة ثالثة تتعلق بالتصفح في نتائج البحث التي تم إرجاعها بواسطة ***PubMed***.

الطريقة المعتادة لتمثيل النص هي حزمة من كلمات ***BoW***. يعاني هذا التمثيل من عدم وجود معنى في التمثيل الناتج ويتجاهل جميع الدلالات الموجودة في النص الأصلي، يتم دمج الدلالات بواسطة تمثيل حزمة من الدلالات ***BoC***. ومع ذلك، من خلال تطبيق هذا التمثيل المفاهيمي، يمكن أن يؤدي تعيين كلمات نصية للمفاهيم المقابلة في المورد الدلالي، إلى غموض في التمثيل النهائي للوثيقة.

للتعامل مع كل من مشكلة تمثيل النص وفهرسة الوثائق الطبية الحيوية، كجزء من هذه الأطروحة، اقترحنا استخدام موارد المصطلح الداخلي في مجال الطب الحيوي. قدمنا أولاً وناقشنا استراتيجيات المفاهيم المختلفة فيما يتعلق بحل الغموض. قدمنا إطارًا عامًا لتصور النص الطبي الحيوي. ثم اقترحنا طريقتين جديدتين لحل غموض المصطلحات الطبية الحيوية - WSD - استنادًا إلى سياق الكلمة الغامضة من خلال استخدام مقاييس التشابه بين مفاهيم التشابه بين مفهومين. بنفس الطريقة، على مستوى فهرسة المستندات الطبية الحيوية، اقترحنا أيضًا استخدام طريقة جديدة تعتمد على التمثيل النظري. يمكن هذا الأخير من دمج وحدة نمطية إلى حل الغموض اللغوي للمصطلحات الغامضة في بنية نظام الفهرسة واسترجاع المعلومات. في هذه الوحدة، عمدنا إلى استغلال الحلول التي اقترحناها بالفعل للتغلب على مشكلة حل الغموض اللغوي لتحسين عملية الفهرسة وبالتالي زيادة أداء أنظمة استرجاع المعلومات المكرسة للحقل الطبي الحيوي.

بالإضافة، تشكل مشكلة التصفح أيضًا تحديًا رئيسيًا آخر للوصول إلى المعلومات في مجال الطب الحيوي. في الواقع، في مرحلة التصفح التي يوفرها محرك ***PubMed*** بالإضافة إلى محركات البحث الأخرى مثل Google و Yahoo و Bing، تقترح إعادة قائمة مرتبة بعشرات من روايات المقتطفات (بيانات التعريف) ، لا يشاهد المستخدمون سوى الصفحات الأولى ، وبالتالي لا يتم الرجوع إلى الوثائق الموجودة في نهاية القائمة على الرغم من أنها قد تكون ذات صلة. لحل مشكلة التصفح هذه، اقترحنا مساهمة أخرى بناءً على استخدام نظام يستند إلى تحليل المفاهيم FCA،

فهي تتيح من ناحية إنتاج مجموعة موضوعية من نتائج البحوث، ومن ناحية أخرى تقديم واجهة تصفح هرمية على مستويان، وبالتالي تمكين تصفح سريع، بسيط وفعال في هيكل شجرة من المجموعات الموضوعية التي أنتجت.

لتسليط الضوء على جميع المساهمات التي اقترحناها في هذه الأطروحة، قمنا بإجراء العديد من دراسات مقارنة. النتائج التي تم الحصول عليها مشجعة للغاية، سواء لغموض المفردات أو لحالة الفهرسة والتصفح.

Résumé

Face à l'immense masse d'informations biomédicale qui ne cesse d'argumenter exponentiellement jours après jours, l'accès à l'information dans le domaine biomédicale devient de plus en plus ardu. Il est donc très intéressant de faire face à cette problématique, de proposer des méthodes, et de mettre en place des outils permettant ainsi d'assister les utilisateurs de l'information dans le domaine biomédical d'y'accéder très rapidement et facilement pour satisfaire le mieux possible leurs besoins d'une manière très pertinente. Afin d'améliorer la performance des résultats de n'importe applications de Fouille de Texte Biomédicale, dans le cadre de cette thèse, nous nous sommes intéressés principalement au problème de la représentation du texte biomédical, ainsi nous traitons les défis de l'indexation des documents biomédicaux. Par ailleurs, nous nous sommes intéressés aussi à un troisième problème lié à la Navigation dans les résultats de recherche retournés par ***PubMed***. La manière habituelle de représenter un texte est un paquet de mots ***BoW***. Cette représentation souffre du manque de sens dans la représentation qui en résulte et ignore toute la sémantique qui réside dans le texte original, l'intégration de la sémantique est résolue par une représentation dite Sac-de-Concepts ***BoC***. Néanmoins, en appliquant cette représentation conceptuelle, l'attribution d'un terme à ses concepts dans une ontologie peut générer de l'ambiguïté dans la représentation finale du document.

Pour faire face à la fois au problème de la Représentation du Texte et à celui de l'Indexation des Documents Biomédicaux, dans le cadre de cette thèse, nous avons proposé d'utiliser des ressources termino-ontologiques du domaine biomédical. En effet, nous avons d'abord présenté et discuter les différentes stratégies de conceptualisation en relation avec la résolution de l'ambiguïté. Ainsi, nous avons présenté un cadre de travail générique pour la conceptualisation de texte biomédical. Ensuite nous avons proposé deux nouvelles méthodes pour la résolution de l'ambiguïté des termes biomédicaux - WSD - basée sur le contexte du mot ambigu à travers l'utilisation des mesures de similarité entre les concepts de l'ontologie utilisée. De même au niveau de l'indexation des documents biomédicaux, nous avons proposé également d'utiliser une nouvelle méthode basée sur la représentation conceptuelle. Cette dernière consiste à intégrer un module pour la désambiguïsation lexicale des termes ambigus dans l'architecture du système d'indexation et recherche d'information. En effet, dans ce module, nous avons exploité encore plus les solutions que nous avons déjà proposées pour surmonter le problème de la désambiguïsation lexicale pour améliorer le processus de l'indexation et par suite augmenter la performance des système de recherche d'information dédiés au domaine biomédical.

Par ailleurs, le problème de Navigation pose aussi un autre défi majeur pour l'accès à l'information dans le domaine biomédical. En effet, dans la phase de la Navigation offerte par le moteur ***PubMed*** ainsi que les autres moteurs de recherche tel que Google, Yahoo, Bing propose de retourner une liste

ordonnée d'une dizaine de milliers de snippets (métas-données), les utilisateurs ne consultent que les premières pages, et par conséquent les documents situés à la fin de la liste ne sont jamais consultés bien qu'ils puissent être pertinents. Pour remédier à ce problème de navigation, nous avons proposé une autre contribution basée sur l'utilisation de l'Analyse Formelle de Concept (Formal Concept Analysis) FCA, cette dernière permet d'une part de produire un regroupement thématique des résultats de recherches, et d'autre part de présenter une interface de navigation hiérarchique sur deux niveaux permettant ainsi une navigation rapide, simple et efficace dans une structure arborescente des groupes thématiques ainsi générés.

Pour mettre en évidence l'intérêt de toutes les contributions que nous avons proposé dans le cadre de ce travail de thèse, nous avons bien mené plusieurs études comparatives. Les résultats obtenus sont très encourageants que ce soit pour la désambiguïsation lexicale ou pour le cas de l'Indexation et la navigation.

Abstract

With the immense mass of biomedical information that continues to argue exponentially days after days, access to information in the biomedical field is becoming increasingly difficult. It is therefore very interesting to deal with this problem, to propose methods, and to set up tools to help users of information in the biomedical field to access it very quickly and easily. In order to improve the performance of the results of any Biomedical Text Mining application, as part of this thesis, we have been interested primarily in the problem of biomedical text representation and we are dealing with the challenges of indexing of biomedical documents. In addition, we were also interested in a third problem related to Browsing in the search results returned by ***PubMed***.

The ordinary way of representing a text is Bag of Words ***BoW***. This representation suffers from the lack of senses in the resulting representation and ignores all the semantics that resides in the original text, the integration of semantics is solved by a Bag of Concepts ***BoC*** representation. Nevertheless, by applying this conceptual representation, assigning a term to its concepts in an ontology can generate ambiguity in the final representation of the document.

To deal with both the problem of Text Representation and Biomedical Document Indexing, we proposed to use termino-ontological resources in the biomedical field. Indeed, we first presented and discussed the different conceptualization strategies in relation to the resolution of ambiguity. Thus, we presented a generic framework for the conceptualization of biomedical text. Then we proposed two new methods for Word Sense Disambiguation -WSD - based on the context of the ambiguous word through the use of similarity measures between the concepts of the used ontology. In the same way, at the indexing level of biomedical documents, we also proposed to use a new method based on conceptual representation. This latter consists in integrating a module for WSD of ambiguous terms into the architecture of the indexing and information retrieval system. Indeed, in this module, we have exploited even more the solutions that we have already proposed to overcome the problem of WSD to improve the process of indexing and consequently increase the performance of information retrieval systems devoted to biomedical field.

In addition, the Navigation problem also poses another major challenge for access to information in the biomedical field. Indeed, in the phase of the Navigation offered by the ***PubMed*** engine as well as other search engines such as Google, Yahoo, Bing proposes to return an ordered list of tens of thousands of snippets (metas-data), users only view the first pages, and therefore the documents at the end of the list are never consulted although they may be relevant. To remedy this navigation problem, we have proposed another contribution based on the use of the Formal Concept Analysis FCA, it allows

on the one hand to produce a thematic group of research results, and on the other hand to present a hierarchical navigation interface on two levels thus enabling a fast, simple and efficient navigation in a tree structure of thematic groups thus generated.

To highlight the interest of all the contributions that we proposed in this thesis, we have conducted several comparative studies. The results obtained are very encouraging, whether for lexical disambiguation or for the case of indexation and navigation.

Liste des abréviations

BioNE	: Biomedical Named Entity	: Entités Nommées Biomédicales
BioNER	: Biomedical Named Entity Recognition	: Reconnaissance des Entités Nommées Biomédicale
BoC	: Bag of Concepts	: Sac de Concept
BoP	: Bag of Phrases	: Sac de Phrases
BoW	: Bag of Word	: Sac de Mots
CRF	: Conditional Random Fields	: Champs Aléatoires Conditionnels
CUI	: Concept Unique Identifier	
DT	: Decision Tree	: Arbre de Décision
FN	: False Negative	: Faux Négatif
FP	: False Positive	: Faux Positif
HMM	: Hidden Markov Model	: Modèle de Markov Caché
IC	: Information Content	: Contenu Informatif
ME	: Maximum Entropy	: Entropie Maximale
MEDLINE	: Medical Literature Analysis and Retrieval System Online	
MEMM	: Maximum Entropy Markov Model	: Modèle de Markov à Entropie Maximale
MeSH	: Medical Subject Headings	
NB	: Naive Bayes	: Naïve Bayésienne
NER	: Named Entity Recognition	: Reconnaissance des Entités Nommées
NLM	: National Library of Medicine	: Bibliothèque Nationale de Médecine
NLP	: Natural Language Processing	: Traitement du Langage Naturel
RST	: Rough Set Theory	: Théorie des Ensembles Approximatifs
SVM	: Support Vector Machine	: Machines à Vecteurs de Support
TP	: True Positive	: Vrai Positif
UMLS	: Unified Medical Language System	: Système de Langage Médical Unifié
VSM	: Vector Space Model	: Modèle vectoriel
WSD	: Word Sense Disambiguation	: Résolution de l'Ambiguïté Lexicale

Liste des Tableaux

Liste des Figures

Table des matières

Chapitre 1 : Contexte et Contributions de la Thèse

1. Contexte du travail

Avec la croissance rapide de la technologie biologique à haut débit, l'étude de la science biomédicale est entrée dans l'ère de l'omique. Celui-ci apporte plusieurs données omiques, notamment la génomique et la transcriptomique ; les énormes quantités de données biologiques continuent à émerger de la recherche en sciences de la vie. Le nouveau phénomène, la nouvelle découverte et les nouvelles données expérimentales dans la recherche biomédicale sont principalement publiés dans des revues scientifiques sous forme de texte électronique. Un grand nombre d'informations biologiques sont dispersées dans toutes sortes d'études. Le traitement de ces littératures biomédicales pourrait extraire davantage d'informations biologiques et découvrir de nouvelles connaissances biomédicales. La recherche manuelle de ces informations est très difficile et non efficace. De ce fait l'exploitation de ces ressources devient impossible.

La littérature biomédicale peut être considérée comme un grand référentiel des données non structurées, qui permet à la fouille de texte d'entrer en jeu. La fouille de texte est apparue comme une solution potentielle pour acquérir des connaissances qui permettent de relier le texte libre à la représentation structurée d'informations biomédicales à l'aide de la technologie d'intelligence artificielle, notamment le traitement du langage naturel, l'apprentissage automatique et la fouille de données (***Data Mining***) pour le traitement des collections de textes volumineux. Par conséquent, la technologie d'extraction de texte est un outil puissant pour extraire des informations précieuses de la littérature biomédicale. Il existe des définitions plus larges de la fouille de texte biomédical. À savoir, tout travail qui extrait des informations à partir de texte pourrait être considéré comme une exploration de texte, qui inclurait une gamme de reconnaissance d'informations statique, allant de l'extraction dynamique d'informations à l'application de la fouille de texte biomédical.

Dans la section suivante nous essayons de donner une introduction à la fouille de textes dans le domaine biomédical en motionnant quelques ressources indispensables dédiés à l'exploration du texte biomédical.

2. Fouille de Texte Biomédical

2.1. Introduction à la Fouille de Texte Biomédical

La fouille de texte ou le texte Mining biomédical est le domaine de recherche qui combine la linguistique informatique, la bio-informatique, la science de l'information médicale, les domaines de recherche, etc. Le développement de la fouille de texte biomédical dure depuis moins de 25 ans [1] et appartient à une branche de la bio-informatique. La bio-informatique est définie comme application de la science de l'information et la technologie qui permet de comprendre,

d'organiser et de gérer des données biomoléculaires. Son objectif est de fournir des outils et des ressources aux chercheurs en biologie, de les aider à obtenir des données biologiques et de les analyser, afin de découvrir de nouvelles connaissances [2] du monde biologique. La fouille de texte biomédical est un sous-domaine de la bio-informatique, il fait référence à l'utilisation de la technologie d'exploration de texte pour traiter des documents de la littérature biomédicale et acquérir des informations biologiques, organiser et gérer les informations biomédicales et les fournir aux chercheurs. Par conséquent, l'exploration biomédicale de textes peut extraire diverses informations biologiques [3], telles que des informations sur les gènes et les protéines, des informations sur la régulation de l'expression des gènes, des informations sur le polymorphisme des gènes et l'épigénétique, ainsi que sur les relations entre gènes et maladies. Les informations biologiques peuvent aider les experts à comprendre les phénomènes de la vie et à comprendre les règles de leurs activités. L'utilisation de ces informations aidera à comprendre le mécanisme de la génération de maladies, à promouvoir le développement de technologies de diagnostic des maladies et aussi, développer de nouveaux médicaments dans le domaine de la recherche biomédicale.

Un grand nombre de méthodes de la fouille de texte ont été établies pour faciliter l'extraction d'informations biologiques. Ces méthodes pourraient être proposées pour extraire : des informations dont le degré de confiance dépend des dictionnaires, des approches basées sur les statistiques et sur les connaissances et de la génération automatique de règles utilisant un étiquetage morpho-syntaxique (Part-Of-Speech - POS) mais aussi de certains algorithmes d'apprentissage automatique.

2.2. Ressources pour fouille du texte biomédical

La principale ressource pour fouille de texte biomédical est évidemment le texte, et cette section présente certaines collections de textes largement utilisés dans le domaine biomédical. Bien que l'exploration de texte n'exige pas l'utilisation de corpus spécialisés ou annotés, les collections annotées manuellement sont souvent plus utiles que les textes originaux. Par exemple, la conception originale de la découverte basée sur la littérature [4] a été facilitée par l'utilisation de MeSH [5] (***Medical Subject Headings***) termes de vocabulaire contrôlé ajoutés aux citations bibliographiques au cours du processus d'indexation MEDLINE [6] (***Medical Literature Analysis and Retrieval System Online***). Avec l'augmentation du nombre de collections annotées accessibles au public, la communauté du traitement du langage biomédical a commencé à se concentrer sur des formats d'annotation, des lignes directrices et des normes interchangeables communes.

A. Corpus

Que ça soit fouille de texte considérée dans le sens strict de la découverte ou dans un sens plus large incluant toutes les étapes de traitement et de recherche du texte menant à la découverte, MEDLINE a été la première et reste la principale ressource de fouille de texte biomédical. La base de données MEDLINE contient des références bibliographiques sur des articles de revues dans les sciences de la vie avec une concentration sur la biomédecine. Elle est gérée par la Bibliothèque national de médecine (National Library of Medicine - NLM) - des États-Unis. MEDLINE 2011 qui contient plus de 18 millions de références publiées de 1946 à ce jour dans plus de 5 500 revues dans le monde. Des résumés de la littérature biomédicale peuvent être obtenus de différentes manières. Pour les besoins de l'exploration de texte, les enregistrements MEDLINE / PubMed peuvent être téléchargés à l'aide des utilitaires de programmation Entrez (E-utilities) [7]. Vous pouvez également obtenir des sous-ensembles de citations MEDLINE dans les archives d'évaluations à l'échelle de la communauté utilisant MEDLINE, ainsi que auprès de groupes de recherche individuels partageant leurs annotations. Ces collections comprennent l'ensemble historique OHSUMED [8] contenant toutes les citations MEDLINE dans 270 revues médicales publiées sur une période de cinq ans (1987-1991) et un ensemble plus récent de données de la piste de génomique de TREC [9] contenant dix années de citations MEDLINE. (1994-2003). Des annotations de remplacement qui prennent en charge la pertinence de la récupération d'informations, la classification de documents et la réponse aux questions sont disponibles pour des parties de ces collections. Alors que les collections TREC donnent accès à des plages MEDLINE sur une période donnée, les autres collections sont orientées tâches. Par exemple, le corpus GENIA [10] contient 1 999 résumés MEDLINE récupérés à l'aide des termes MeSH «humain», «cellules sanguines» et «facteurs de transcription». Le corpus GENIA est actuellement la collection de résumés MEDLINE la plus complètement annotée. Il est annoté pour une partie du discours, la syntaxe, la coréférence, les concepts et événements biomédicaux, la localisation cellulaire, les associations maladie-gène et les voies d'accès. En outre, le corpus GENIA est l'un des trois constituants du corpus BioScope [11], qui fournit des résumés GENIA MEDLINE, cinq articles en texte intégral et une collection de rapports de radiologie annotés avec des indications de négation et de modalité, ainsi que leur portée. Parmi les autres collections de résumés MEDLINE annotées localement, citons les collections précédentes de BioCreAtIve [12] [13] et le corpus PennBioIE [14] [15]. Le corpus PennBioIE contient 1100 résumés pour les enzymes du cytochrome P-450 et 1157 résumés en oncologie avec des annotations pour les paragraphes, les phrases, les jetons, les parties du discours, la syntaxe et les entités biomédicales. Enfin, l'initiative Collaborative d'Annotation d'un

grand corpus biomédical (CALBC) [16] a proposé la création d'un corpus «Silver Standard» contenant des résumés MEDLINE automatiquement annotés avec des entités biomédicales par les participants. Ce corpus vient tout juste d'être rendu public.

Informatifs et sans aucun doute utiles pour l'exploration de texte, les résumés MEDLINE ne contiennent pas toutes les informations présentées dans des articles en texte intégral. Certaines informations (par exemple, les réglages exacts d'une expérience ou la discussion des résultats) sont presque exclusivement contenues dans le corps d'un article. La promesse d'une augmentation qualitative de la quantité d'informations utiles a entraîné la création de plusieurs collections de textes intégraux. Par exemple, le corpus TREC Genomics Track contient environ 160 000 articles en texte intégral d'environ 49 revues sur la génomique, qui ont été obtenus au format HTML à partir de la distribution électronique des revues de Highwire Press [17]. Une autre collection d'articles en texte intégral annotés et pertinents pour les descriptions de cas de patients a été développée dans les évaluations d'ImageCLEF [18] [19]. Le corpus de texte intégral du Colorado richement annoté [20] ajoute au corpus croissant de collections des textes intégraux sémantiquement et syntaxiquement annotés (y compris la partie de la collection BioScope en texte intégral). Enfin, la plus grande source d'articles en texte intégral originaux disponibles au public est le sous-ensemble Open Access de PubMed Central [21].

Avec l'intérêt croissant pour l'extraction de textes cliniques et la bio-surveillance, plusieurs collections publiques de textes cliniques sont devenues disponibles. Ces collections incluent des rapports dans la base de données MIMIC II (MIMIC II) [22], la collection de rapports cliniques de Pittsburgh [23] et les collections annotées i2b2 [24] [25] [26] [27]. Plusieurs études récentes ont utilisé le Web (c'est-à-dire Twitter et les blogs et sites communautaires liés à la santé) comme un corpus, mais il n'est pas clairement défini si les collections créées pour ces études sont ou non accessibles au public.

B. Annotation

L'annotation de texte biomédical ajoute des informations à une collection de documents qui peuvent ensuite être exploitées à des fins de fouille de texte. En général, les annotations de documents dans le domaine biomédical suivent les principes énoncés dans le traitement de langage naturel en domaine ouvert en ajoutant des annotations à plusieurs niveaux d'analyse linguistique. Les différents aspects impliquent des annotations grammaticales (y compris morphologiques et syntaxiques), sémantiques et pragmatiques [28]. La grammaire et la sémantique sont si étroitement liées que la plupart des efforts d'annotation combinent les deux. Par exemple, les créateurs de corpus pourraient décider d'annoter les entités nommées d'intérêt uniquement dans les expressions

nominales. Comme alternative, Wilbur et *al.* [29] se concentrent sur l'annotation des «fragments contenant de l'information dans un texte scientifique» sans spécifier de restrictions grammaticales. Les auteurs définissent les cinq axes d'annotation suivants : mise au point, polarité, certitude, preuve et direction. Ces classes sont principalement utilisées au niveau de la phrase. Ces phrases peuvent être segmentées si nécessaire au cas où un changement dans l'un des aspects de l'annotation est détecté. Cependant, même les annotations centrées sur le sens ne peuvent être complètement exemptes de grammaire. Par exemple, l'un des indices permettant d'annoter des fragments en tant que Evidence est un verbe au passé indiquant une observation ou une constatation. Les guides publiés par les auteurs [30] constituent un bon point de départ pour développer d'autres directives d'annotation en fouille de texte dans le domaine biomédical. Il existe trois approches pour l'annotation d'un texte biomédical.

Ces méthodes incluent :

1. Une annotation manuelle complète basée sur les connaissances des annonceurs ;
2. Une annotation assistée, dans laquelle la sortie d'un outil d'annotation est corrigée manuellement ;
3. Une annotation basée sur une ontologie (manuelle ou assistée) et dans laquelle seuls les termes et relations présents dans une source de connaissance existante sont annotés.

Chacune de ces approches a ses forces et ses faiblesses. Par exemple, une annotation assistée est généralement plus cohérente, mais elle peut être biaisée. De même, une annotation basée sur une ontologie sera probablement biaisée au profit de faits connus. Avoir plus d'un annotateur pour chaque document texte et plusieurs groupes d'annotateurs peuvent compenser de tels biais [31]. Outre, les outils d'extraction d'informations génériques pouvant être utilisés pour faciliter l'annotation (décrits ci-dessous), plusieurs outils d'exploration de texte ont été développés pour prendre en charge spécifiquement le processus d'annotation. Knowtator [32] et eHOST (Extensible Human Oracle Suite of Tools) [33] sont des exemples d'outils largement utilisés pour l'annotation de texte biomédical, le dernier étant de plus en plus utilisé pour l'annotation de texte clinique. Pour que ces outils soient utiles, ils doivent être faciles à utiliser, prendre en charge divers types d'annotation et permettre une annotation collaborative, entre autres facteurs [34] [35].

C. Ressources de Connaissances Externes

Le domaine biomédical offre un riche ensemble de ressources de connaissances externes prenant en charge les applications de fouille de texte. Le système de langage médical unifié – Unified Medical Language System (UMLS) [36], un recueil de vocabulaires contrôlés géré par la NLM, est la ressource la plus complète, regroupant plus de 100 dictionnaires, terminologies et

ontologies dans son métathesaurus. Il fournit également un réseau sémantique qui représente les relations entre les entrées de Metathesaurus, un lexique contenant des informations lexicographiques sur les termes biomédicaux et les mots anglais courants, ainsi qu'un ensemble d'outils lexicaux. Dans l'ensemble, la NLM fournit plus de 200 sources de connaissances et outils pouvant être utilisés pour la fouille de texte [37]. D'autres ensembles d'ontologies sont gérés de manière collaborative par la fonderie OBO [38] et le Centre National d'Ontologie Biomédicale (NCBO) [39]. Les ontologies NCBO sont accessibles et partagées via BioPortal [40]. Parmi les autres grands centres qui gèrent des ressources spécialisées pour la fouille de texte biomédical, citons « British National Centre [41] » et « European Bioinformatics Institute » [42]. En plus de ces ressources étendues, le domaine biomédical offre des sources de connaissances approfondies axées sur des sous-domaines spécifiques de la biomédecine. Par exemple, la « Pharmacogenomics Knowledge Base » [43] est une collection de publications scientifiques annotées avec des données de génotype primaire et de phénotype, des variantes de gènes et des relations gène-médicament-maladie. Les annotations sont téléchargeables à des fins de recherches individuelles. Une autre source spécialisée, « Neuroscience Information Framework [44] », comprend une ontologie couvrant l'anatomie du cerveau, les cellules, les organismes, les maladies, les techniques et d'autres domaines de la neuroscience. La meilleure source de connaissances pour une tâche d'extraction de texte donnée est déterminée par la nature du problème à résoudre. Par exemple, pour explorer la littérature scientifique concernant les relations entre gènes, maladies et médicaments, il faut d'abord reconnaître les exemples de ces entités. Pour faciliter cette tâche, un chercheur peut s'appuyer sur la connaissance des types sémantiques correspondants dans l'UMLS ou choisir d'utiliser des sources de connaissances individuelles, telles que Gene Ontology [45], SNOMED Clinical Terms [46] ou médicaments approuvés par la FDA avec des évaluations d'équivalence thérapeutique (Livre Orange) [47]. Les approches des différentes tâches de fouille de texte dans le domaine biomédical utilisent largement les ressources décrites dans cette section et dérivent parfois des méta-ressources pour une tâche spécifique. Par exemple, Rinaldi et *al.* [48] définissent plusieurs types d'entités nécessaires à l'exploration de la littérature pour les interactions entre protéines (noms de protéines / gènes, composés chimiques, lignées cellulaires, etc.), puis agrègent automatiquement les termes extraits de ressources organisées telles que les identifiants UMLS et Affymetrix pour biopuces multiples, bases de données sur les organismes, et autres, dans une liste de 2 347 734 termes.

D. Outils supportées

La variété et la finalité des outils supportant fouille de texte biomédical font appel à celle des ressources de connaissances décrites ci-dessus. La discussion suivante sur les outils d'extraction de texte omet les applications décrites dans les enquêtes récentes et se concentre plutôt sur les outils de base largement utilisés pour identifier les entités nommées et les relations, ainsi que sur les plates-formes permettant de créer des pipelines d'extraction de texte. MetaMap [49] est l'outil le plus utilisé pour la reconnaissance d'entités nommées basé sur l'UMLS. MetaMap est une application hautement configurable qui identifie les concepts du Meta-thésaurus UMLS en texte libre. Etant donné que MetaMap fournit un large éventail d'options de configuration et s'appuie sur le métathesaurus UMLS, il n'est pas facile de déterminer la meilleure configuration pour une tâche donnée. Cependant, l'exploration des options à l'aide du site Web interactif de MetaMap peut être utile dans ce cas. MetaMap, qui était fourni sous forme de service jusqu'à récemment, est maintenant libre et disponible au téléchargement. ABNER [50] et BANNER [51] sont deux outils statistiques largement utilisés pour la reconnaissance d'entités nommées biologiques. ABNER et BANNER reposent tous deux sur des champs aléatoires conditionnels (Conditional Random Fields - CRF), ainsi qui reposent sur une large fonctionnalité. Contrairement à ABNER, BANNER évite les fonctionnalités sémantiques, mais utilise des fonctionnalités syntaxiques. Les deux systèmes exploitent des caractéristiques linguistiques spécifiques à un domaine telles que la capitalisation, les formes de mots, les préfixes, les suffixes et les lettres grecques. Les outils d'extraction de relations ne sont pas encore aussi facilement accessibles que les outils de reconnaissance d'entités. Kabiljo et *al.* [52] ont comparé les outils disponibles pour l'identification de relations biomédicales (AkanePPI, Whatizit et OpenDMAP) à une approche simple, basée sur des expressions régulières, et ont révélé que l'approche simple fonctionnait étonnamment bien. Les auteurs concluent qu'il est possible d'obtenir un rappel élevé (environ 90%) pour extraire les relations entre gènes et protéines lorsque les outils disponibles sont combinés. Une tendance récente dans le développement et l'utilisation d'outils est l'assemblage de pipelines basé sur des frameworks open-source, tels que l'architecture généralisée pour l'ingénierie du texte (GATE) [53] et l'architecture non structurée de gestion de l'information (UIMA) [54]. MedLEE [55] est le système de traitement de texte clinique le plus abouti (allant de l'identification des problèmes des patients aux événements). Des descriptions d'autres systèmes et tâches d'exploration de texte clinique sont disponibles dans une revue récente [56].

Cette section n'a présenté qu'une partie des ressources de fouille de texte biomédical comme domaine ouvert. Pour compenser les progrès rapides de la recherche liée à la fouille de texte biomédical, de nombreux chercheurs gèrent des sites Web contenant des liens vers des ressources

utilecs (par exemple, BioNLP [57]). En apercevant que cette tâche prend trop de temps à des chercheurs individuels, le département américain des Anciens Combattants et la NLM fournissent un registre d'outils d'extraction de textes biomédicaux, connu sous le nom d'ORBIT, qui est géré par la communauté des chercheurs [58].

3. Objectif de la thèse et principales contributions

3.1. Objectif et motivation

Les travaux présentés dans cette thèse se situent dans le contexte précis de l'accès à l'information médicale. Plus précisément, nous nous sommes intéressés principalement aux problèmes suivants :

Reconnaissance d'Entités Nommées Biomédicales [Bio Named ER] : Les entités nommées biomédicales sont des expressions ou des combinaisons d'expressions qui désignent des objets ou des groupes d'objets spécifiques dans la littérature biomédicale. Ainsi, la reconnaissance des entités nommées est une tâche importante dans le traitement automatique des articles scientifiques biomédicaux. La reconnaissance des entités nommées biomédicale (BioNER) est la tâche qui vise à reconnaître automatiquement les entités nommées biomédicales telles que les gènes, les protéines, les cellules, les médicaments, les maladies, etc. BioNER joue un rôle vital et peut avoir un impact positif ou négatif sur toutes les applications de traitement intelligent du texte biomédical, telles que la catégorisation, le regroupement, les systèmes question/réponse à un texte biomédical, la recherche d'information biomédicale et la création d'ontologies biomédicales. De nombreux systèmes de BioNER ont été proposés dans la littérature [59]. Par conséquent, il sera très utile d'évaluer les performances d'un tel système de BioNER.

Intégration de la sémantique dans la représentation du texte biomédical et la résolution de l'ambigüité lexicale des termes biomédicaux : La manière habituelle de représenter un texte est un paquet de mots *BoW*. Cette représentation souffre du manque de sens dans la représentation résultante qui ignorent toute la sémantique qui réside dans le texte original ; L'intégration de la sémantique dans la représentation textuelle est résolue par l'intégration des concepts - en tant qu'unité de connaissance. Cette intégration est due à un processus de conceptualisation qui est par définition un mappage des mots de texte aux concepts correspondants dans la ressource sémantique. Lors de la recherche dans la ressource sémantique pour le mappage d'un mot polysémique, la conceptualisation peut trouver plusieurs correspondances ayant des significations différentes. Par exemple, le mot "Book" signifie en anglais un livre et une réservation (Ticket, hébergement, etc.). Pour faire face à ce problème, la résolution de l'ambiguïté lexicale identifie automatiquement le sens approprié d'un mot ambigu en fonction du contexte dans lequel le mot est utilisé.

Indexation des documents biomédicaux : L'indexation des documents consiste à extraire, organiser, structurer, indexer, sauvegarder le contenu sémantique des documents dans des structures de données appelées index. L'indexation permet de retrouver rapidement les documents contenant un ou plusieurs termes donnés et éventuellement ses positions, ses fréquences d'occurrence dans un document. Nous distinguons deux types d'indexation : Indexation libre et Indexation contrôlée. Dans l'indexation libre, l'indexeur extrait les mots-clés d'un document ou les choisit librement sans l'aide de sources de connaissance. Alors que dans l'indexation contrôlée : le vocabulaire ou le langage d'indexation est déjà défini a priori et son utilisation exclusive s'impose à l'indexeur. Notons que ces deux types d'indexation peuvent se réaliser manuellement ou automatiquement. L'indexation contrôlé automatique ou indexation sémantique est plus robuste vue qu'elle se base sur des ressources de connaissances terminologiques. Cependant, l'hors de l'utilisation des ressources de connaissance du domaine biomédical, il présente des problèmes généralement lié à l'ambigüité des termes biomédicaux.

Navigations et de représentations des résultats de recherche des documents biomédicaux : Le processus de navigation des résultats de recherche est l'un des problèmes majeurs des moteurs de recherche Web traditionnels. PubMed est un moteur de recherche très populaire dans le domaine des sciences de la vie et de la recherche biomédicale et il permet un accès gratuit aux bases de données accédant principalement à la base de données MEDLINE contenant des références et des résumés sur les sciences de la vie et les sujets biomédicaux. Comme tous les autres moteurs de recherche, PubMed a pour objectif de récupérer les citations jugées pertinentes pour une requête utilisateur. Les développeurs de moteurs de recherche modernes ont consacré beaucoup d'efforts à l'optimisation du classement des résultats de recherche, dans l'espoir de placer les plus pertinents en haut de la liste. Néanmoins, aucune solution de classement n'est parfaite, en raison de la complexité inhérente au classement des résultats de recherche.

3.2. Principales contributions

Dans le cadre de cette thèse, nos contributions portent sur les quatre volets présentés dans la sous-section précédente.

Pour remédier au problème de la reconnaissance des entités nommées biomédicale, nous avons réalisé une étude comparative d'une part théorique, et d'autre part expérimentale entre sept méthodes basées sur l'approche d'apprentissage automatique. Cette étude très riche a été couronnée par un papier « ***A Comparative Study of Biomedical Named Entity Recognition Methods Based Machine Learning Approach*** ».

Pour remédier au problème de la désambiguïsation des termes biomédicaux, nous avons réalisé une étude sur l'impact de la distance des termes du contexte sur l'algorithme **SenseRelate**. Pour y parvenir, nous avons apporté quelques modifications à ***SenseRelate*** d'une manière que le score final du terme en cours d'analyse ne sera pas basé sur la distance entre lui-même et les termes de son contexte, ce qui est peut être considérée comme une nouvelle méthode que nous avons nommés **NoDistanceSenseRelate**. De plus, nous avons effectué une étude comparative entre les deux algorithmes (**SenseRelate** et **NoDistanceSenseRelate**), ce travail a été publié avec un papier. « ***Biomedical Word Sense Disambiguation Context-Based: Improvement of SenseRelate Method*** ».

Nous avons ainsi proposé une nouvelle mesure de similarité pour le texte court basée sur la théorie des ensembles approximatifs. Cette mesure a été intégrée dans un algorithme ***Web-kernel*** de désambiguïsation des termes biomédicaux en utilisant ***PubMed*** comme étant une grande ressource d'information biomédicale. Cette étude a été finalisé par un article « ***An Efficient Method for Biomedical Word Sense Disambiguation Based On New Web-Kernel Similarity*** ».

Pour remédier au problème de représentation des documents du domaine biomédical, nous avons proposé d'utiliser une méthode basée sur l'approche conceptuelle qui consiste à représenter un document par un vecteur de concepts (***Bag of Concepts - BoC***) au lieu et à la place d'un vecteur de mots (***Bag of Words - BoW).*** Cette dernière méthode, consiste à intégrer un block pour la désambiguïsation lexicale des termes ambigus.

Dans l'objectif d'évaluer la solution proposée, nous avons commencé bowd'abord par la réalisation d'un système de catégorisation des documents biomédicaux en intégrant ***SenseRelate/NoDistanceSenseRelate*** dans le processus de représentation du texte biomédical, de plus, nous avons bien mené une étude comparative entres les trois stratégies de désambiguïsation ***First Concept, All Concept, Context-Based*** à travers une série d'expérimentations en se basant sur l'ontologie du domaine biomédical UMLS. Cette étude a été couronnée par un papier « ***Evaluation of Disambiguation Strategies on Biomedical Text Categorization*** ».

Pour remédier au problème d'indexation, nous avons aussi intégré notre solution proposée de représentation conceptuelle basée sur la désambiguïsation lexicale des termes biomédicaux en utilisant les deux méthodes ***SenseRelate*** et ***NoDistanceSenseRelate*** dans un processus d'indexation sémantique des documents biomédicaux. De plus, nous avons bien mené une étude évaluative du système de recherche d'information basé sur l'indexation sémantique proposé à travers une série d'expérimentations sur 63 requêtes (chacune est fournie avec un ensemble de documents jugés pertinents par un groupe de médecins). Cette étude a été publiée comme article «*An Empirical Study of Word Sense Disambiguation for Biomedical Information Retrieval System*».

Pour le problème de navigation, nous avons effectué une étude du problème de représentions des documents biomédicaux dans le moteur de recherche ***PubMed*** pour faciliter l'accès à l'information biomédical, ensuite nous avons proposés, un système basé sur l'analyse formelle de concepts (FCA) pour le regroupement des résultats de recherche d'information biomédicale en fonction de leur structure hiérarchique et nous avons évalué ses performances. Cette étude a été finalisée par un article « ***Formal Concept Analysis for Biomedical Web Search Results Clustering*** »

4. Organisation de la thèse

4.1. Présentation générale

Le présent rapport est composé de deux parties, précédées par un chapitre présentant le contexte général de la thèse et une introduction générale et suivies d'une conclusion et des perspectives. La première partie intitulée Représentation et Conceptualisation du Texte Biomédical, composée de 3 chapitres. Chapitre 2, présente une étude concernant la reconnaissance des entités nommées biomédicales, chapitre 3 traite la résolution de l'ambigüité dans la Représentation Conceptuelle, et chapitre 4, présente l'Enrichissement Sémantique du texte court : Web-Kernel. La deuxième partie dénommée Applications : Catégorisation, Indexation/Recherche d'Information et Navigation (PubMed) contient les chapitres 5, 6 et 7. Dans cette partie nous allons présenter deux études d'évaluation de la représentation proposée dans la partie 1 par deux applications de fouille de texte biomédical : La Catégorisation du Texte dans le chapitre 5 et l'Indexation/Recherche d'information dans le chapitre 6. Par ailleurs, dans la même partie, le chapitre 7 concerne une étude du problème de navigation (représentions des documents biomédicaux).

4.2. Contenu des chapitres

Chapitre 2 : Reconnaissance des Entités Nommées Biomédicale

Dans ce chapitre, nous présentons une étude du système de reconnaissance des entités nommées Biomédicale avec une étude comparative.

Chapitre 3 : La Résolution de l'Ambigüité dans la Représentation Conceptuelle.

Dans ce chapitre nous allons discuter des différentes stratégies de conceptualisation et de résolution de l'ambiguïté et présenté un cadre de travail générique pour la conceptualisation de texte biomédical. Nous allons également présenter une étude sur les mesures de similarité entre deux concepts pour la résolution de l'ambiguïté lexicale des termes biomédicaux.

Chapitre 4 : Enrichissement Sémantique du texte court : Web-Kernel.

Dans ce chapitre nous présentons une nouvelle mesure de similarité ***Web-kernel*** qui se base sur l'enrichissement Sémantique du texte court.

Chapitre 5 : Catégorisation du Texte Biomédical.

Dans ce chapitre, nous étudions l'impact des trois stratégies de désambiguïsation sur la catégorisation de texte et nous évaluons la contribution des algorithmes de WSD à la catégorisation de texte biomédical en utilisant trois modèles d'apprentissage automatique : Machine à vecteur de support (SVM), Naïve Bayes (NB) et Entropie Maximum (ME).

Chapitre 6 : Indexation et Recherche d'Information Biomédicale.

Dans ce chapitre nous présentons une étude sur l'impact des stratégies de conceptualisation tout en appliquant les trois stratégies de désambiguïsation au processus d'indexation et nous évaluons la contribution des algorithmes de désambiguïsation de sens (WSD) au système de recherche d'informations biomédicales.

Chapitre 7 : Navigation et Représentations des Résultats de Recherche de PubMed.

Dans ce chapitre nous abordons le problème de navigation (représentions des documents biomédicaux), dont nous présentons et nous évaluons une nouvelle méthode pour le regroupement des résultats de recherche biomédicale sur ***PubMed***, basée sur leur structure hiérarchique en utilisant à l'Analyse Formelle de Concept - FCA.

Partie I : Représentation et Conceptualisation du Texte Biomédical

- **Chapitre 2 : Reconnaissance des Entités Nommées Biomédicales**
- **Chapitre 3 : Résolution de l'Ambigüité dans la Représentation Conceptuelle**
- **Chapitre 4 : Enrichissement Sémantique du Texte Court : Web-Kernel**

Introduction de la partie I

La représentation des documents est une étape cruciale pour extraire des modèles d'index ou implémenter les applications de fouille de texte. Pour appliquer diverses techniques d'apprentissage automatique et de fouille de texte, il est nécessaire de transformer les documents bruts en vecteurs numériques, dans lesquels les caractéristiques qui définissent chaque document sont capturées. Si ces vecteurs de documents peuvent conserver une proximité appropriée entre les documents et leurs caractéristiques uniques, les algorithmes de feuille de texte peuvent utiliser des informations plus précises qui sont masquées dans les données.

Deux méthodes de représentation de documents biomédicaux sont principalement utilisées pour résoudre les problèmes de fouille de de texte. Reconnue pour son interopérabilité simple et intuitive, la méthode du Sac de Mots (***Bag of Word - BoW***) représente un vecteur de document par ses mots fréquents. D'autre part, la représentation du texte biomédical en tant que Sac de Concept (***Bag of Concepts - BoC***) utilisant des Ressources de Connaissances permet d'extraire les sens et la sémantique qui résident dans le texte original.

La plupart des méthodes de représentation de documents les plus répandues reposent souvent sur les approches basées sur le Sac de Mots [60] [61], selon lesquelles un document est fondamentalement représenté par le nombre d'occurrences de mots qu'il contient. Pendant des décennies, cette approche s'est révélée efficace pour diverses tâches d'extraction de texte [62] [63] [64]. L'un de ses principaux avantages est qu'il génère des vecteurs de documents pouvant être interprétés de manière intuitive. Chaque caractéristique d'un vecteur de document indique l'occurrence d'un mot spécifique dans un document. L'approche du ***BoW*** peut toutefois être problématique lorsque le nombre de documents représentés est très grand, ce qui est le cas dans le domaine biomédical. Au fur et à mesure que plusieurs documents augmentent, le nombre de mots uniques dans l'ensemble de documents augmente également. En conséquence, les dimensions des vecteurs de documents générés seront très grandes. Ainsi, cette représentation souffre du manque de sens dans la représentation résultante qui ignorent toute la sémantique qui réside dans le texte original ; D'autre part, la représentation d'un texte biomédical en tant que Sac de Concept utilisant des ressources de connaissances externes permet d'extraire les sens qui résident dans le texte original à travers les concepts des ressources de connaissance externes et la sémantique qui résident dans le texte original.

L'intégration de la sémantique dans la représentation textuelle est résolu par l'utilisation des concepts - en tant qu'unité de connaissance, cette intégration est due à un processus de conceptualisation, ce derniers consiste à effectuer une étape de mappage des mots de texte aux concepts correspondants dans la ressource sémantique. Lors de la recherche dans la ressource sémantique

pour le mappage d'un mot polysémique, la conceptualisation peut trouver plusieurs correspondances ayant des significations différentes. Par exemple, le mot "Book" signifie en anglais un livre et une réservation (Ticket, hébergement, etc.). Pour faire face à ce problème, la Résolution de l'Ambiguïté Lexicale (***Word Sense Disambiguation - WSD***) identifie automatiquement le sens approprié d'un mot ambigu en fonction du contexte dans lequel le mot est utilisé.

Pour la désambiguïsation des sens basée sur la connaissance non supervisée, afin de déterminer le sens correct (concept correct), on utilise généralement des mesures de similarité sémantique ou de connexité [65] qui tentent de quantifier la proximité sémantique entre deux concepts.

En générale, la représentation des documents passe par une étape de Reconnaissance des Entités Nommées (***Named Entity Recognition - NER***) présentes dans le texte. L'objectif de cette étape est d'identifier les termes du lexique biomédical.

Dans cette partie qui concerne une étude sur la Représentation et Conceptualisation du texte biomédical, nous commençons par une étude du système de Reconnaissance des Entités Nommées Biomédicale (***Biomedical Named Entity Recognition – BioNER***), dont nous allons présenter une étude théorique et expérimentale comparative entre sept méthodes d'apprentissage automatique pour ***BioNER***. Nous allons également discuter les différentes stratégies de conceptualisation et de résolution de l'ambiguïté et présenté un cadre de travail générique pour la conceptualisation de texte biomédical. Par la suite, nous présentons deux études en relation avec les mesures de similarité entre deux concepts. La première est une évaluation d'une méthode nommée ***SenseRelate*** utilisée pour déterminer le concept de mot ambigu le plus approprié au contexte ; cette méthode est basée principalement sur l'utilisation d'une fenêtre dite de contexte ; nous évaluons également l'effet de la taille de la fenêtre de contexte sur cette méthode, en utilisant le degré de similarité sémantique et de connexité entre les concepts possibles et les termes entourant le mot ambigu. Dans la deuxième étude nous proposons une mesure de similarité au niveau du noyau Web (***Web-kernel***) entre une paire de textes courts biomédical, dérivés de la définition de concepts et le résultat de recherche du moteur de recherche ***PubMed*** basés sur la théorie des ensembles approximatifs (***Rough Set Theory - RST***).

Cette partie est constituée de trois chapitres :

- **Chapitre II** : Reconnaissance des Entités Nommées Biomédicales.
- **Chapitre III** : Résolution de l'Ambigüité dans la Représentation Conceptuelle.
- **Chapitre IV :** Enrichissement Sémantique : Web Kernel.

Chapitre 2 : Reconnaissance des Entités Nommées Biomédicales

1. Introduction

Deux méthodes de représentation de documents biomédicaux sont principalement utilisées pour résoudre les problèmes de fouille de textes biomédicaux. Reconnue pour son interopérabilité simple et intuitive, la méthode du ***BoW*** représente un vecteur de document par ses mots clé dont on a besoin d'extraire du texte source tous ces termes biomédicaux. D'autre part, la représentation d'un texte biomédical en tant que ***BoC*** utilise des ressources de connaissances externes permet d'extraire les sens qui résident dans le texte original à travers les concepts des ressources de connaissance externes et la sémantique qui résident dans le texte original.

La représentation des documents passe par une étape de reconnaissance des entités nommées présentes dans le texte. L'objectif de cette étape est d'identifier les termes du lexique biomédical.

Les entités nommées biomédicales (***Biomedical Named Entity - BioNE***) sont des expressions ou des combinaisons d'expressions qui désignent des objets ou des groupes d'objets spécifiques dans la littérature biomédicale. En effet, La Reconnaissance des entités nommées (***NER***) est une tâche importante dans le traitement automatique des documents biomédicaux. La reconnaissance des entités nommées biomédicale (***BioNER***) est la tâche qui vise à reconnaître automatiquement les entités nommées biomédicales telles que les noms des gènes, des protéines, des cellules, des médicaments, des maladies, etc.

BioNER joue un rôle vital et peut avoir un impact positif ou négatif sur toutes les applications de traitement intelligent du texte biomédical, telles que la catégorisation, le regroupement, les systèmes question/réponse à un texte biomédical, la recherche d'information biomédicale et la création d'ontologies biomédicales. De nombreux systèmes de ***BioNER*** ont été proposés dans la littérature [59]. Par conséquent, une étude comparative sera d'une très grande importance pour le choix d'un tel système de ***BioNER***.

Nous pouvons diviser les méthodes de ***BioNER*** en quatre approches qui sont basée sur : la recherche dans dictionnaire, les règles linguistique, l'apprentissage automatique et des mesures statistiques. Les méthodes basée sur l'approche d'apprentissage automatique sont plus efficaces que celles d'autres approches et ont été largement utilisées pour apprendre à reconnaître les ***BioNE***s.

Nous proposons une méthode de ***BioNER*** basée sur les modèles d'apprentissage automatique, pour choisir le meilleur model, dans le cadre de ce chapitre nous étudions dans un premier temps le système de ***BioNER***, ensuite nous proposons de réaliser une étude théorique et expérimentale comparative entre sept différentes méthodes d'apprentissage automatique pour ***BioNER*** : Les Champs Aléatoires Conditionnels (***Conditional Random Fields – CRF***), Modèle de Markov

Caché (***Hidden Markov Model – HMM***), Modèle de Markov à Entropie Maximale (***Maximum Entropy Markov Model – MEMM***), Les Machines à Vecteurs de Support ou Séparateurs à Vaste Marge (***Support Vector Machine – SVM)***, Naïve Bayésienne (***Naive Bayes – NB***), L'Entropie Maximale (***Maximum Entropy – ME***) et Arbre de Décision (***Decision Tree – DT***).

La suite de ce chapitre est organisée comme suit : nous présentons d'abord un aperçu des approches de BioNER. Par la suite, nous poursuivons en mettant en évidence les sept méthodes d'apprentissage automatique utilisées et présentons les avantages et les inconvénients de ces modèles. Nous présentons ensuite les corpus utilisés, la méthodologie d'évaluation et les résultats du système d'évaluation, vers la fin du chapitre nous concluons notre étude.

2. Approches de BioNER

Dans cette sous-section nous présentons en détails les quatre approches de BioNER qui sont basée sur : la recherche dans un dictionnaire, les règles linguistique, l'apprentissage automatique et des mesures statistiques.

2.1. Approche basée sur la recherche dans un dictionnaire

Les approches basées sur la recherche dans un dictionnaire identifient les NEs dans un texte en recherchant (par exemple, une correspondance de chaîne) une liste fournie de NEs connus, généralement extraite à partir de lexiques ou de bases de données existantes. La seule hypothèse est l'existence d'une telle liste. Les efforts représentatifs incluent Hirschman et al [66], qui ont eu recours à une stratégie de correspondance de modèle simple pour localiser des gènes à l'aide de noms de gènes dans Fly Base [67]. Egorov et al [68] ont proposé une approche simple et pratique basée sur un dictionnaire pour la reconnaissance des protéines dans les résumés de MEDLINE. Les résultats ont montré que l'approche basée sur le dictionnaire, si elle était utilisée seule, n'atteindrait pas une performance satisfaisante, et qu'il faudrait donc déployer davantage d'efforts pour utiliser cette approche en combinaison avec des approches basées sur des règles et / ou un apprentissage automatique.

2.2. Approche basée sur les règles linguistique

Les approches basées sur des règles linguistique tentent généralement de reconnaître les bio-NE par des règles prédéfinies. Les règles décrivent généralement des structures de dénomination communes en utilisant des indices orthographiques ou lexicaux, ou des caractéristiques morpho-syntaxiques. L'élaboration de telles règles nécessite généralement une connaissance approfondie

de la linguistique, de la biologie, de la médecine et éventuellement des langages de programmation. Ananiadou et *al.* [69] ont mis en œuvre une grammaire et un lexique morphologiques calculés, et l'applique à la reconnaissance de termes médicaux. Fukuda et *al.* [70] ont présenté un système KEX permettant de reconnaître les noms de protéines en utilisant des indices de surface sur des chaînes de caractères. Proux et *al.* [71] ont utilisé des informations lexicales et morphologiques pour détecter des noms de gènes dans des documents biomédicaux. La force de l'approche basée sur des règles linguistique réside dans le fait que les règles peuvent être soigneusement conçues pour traiter des phénomènes linguistiques spécifiques. Cependant, cette approche présente certains inconvénients : l'exigence d'une connaissance approfondie du domaine pour l'élaboration d'une bonne règle. De plus, ces règles sont définies pour un domaine particulier, ce qui rend difficile son application à d'autres domaines, ce qui prend beaucoup de temps.

2.3. Approche basée sur des mesures statistiques

Les approches basées sur les mesures statistiques ont été proposées pour identifier des termes techniques désignant les EN. Frantzi et *al.* [72] ont proposé une méthode appelée C / NC-value, pour extraire automatiquement les termes et les entités nommées dans plusieurs domaines à part le domaine biomédical. Cette méthode a été utilisée plus tard pour Bio-NER. Hliaoutakis et *al.* [73] ont également utilisé la valeur C / NC pour reconnaître les entités nommées biomédicales désignées dans la littérature. Bechet et *al.* [74] ont proposé un système basé sur le modèle d'arbre de décision, permettant d'extraire et d'assigner des tags des entités nommées inconnues. En général, les méthodes statistiques présentent des performances insatisfaisantes et cèdent la place à une approche basée sur l'apprentissage automatique.

2.4. Approche basée sur l'apprentissage automatique

La plupart des méthodes d'apprentissage automatique utilisées dans ***BioNER*** utilisent des données d'apprentissage pour acquérir des caractéristiques utiles à la détection des contours d'EN et à la classification de types d'EN. Le principal problème de cette approche est qu'elle nécessite des ressources d'apprentissage fiables. Toutefois, une bonne partie des recherches actuelle sont centrée sur les ***HMM***, les ***MEMM*** et, récemment, les ***CRF***. La première tentative d'utilisation du HMM a été réalisée par Collier et *al.* [75], qui ont formé un ***HMM*** linéaire interpolation avec des bi-grammes, pour identifier des gènes et des produits géniques. Andrew McCallum [76] a proposé une alternative aux ***HMM*** et ***MEMM***. Settles [77] a construit un système de NER en utilisant des ***CRF***s, qui pourraient reconnaître plusieurs classes d'entités au même moment.

En général, les performances des méthodes basée sur l'apprentissage automatique sont meilleures que celles des approches basées sur des dictionnaires et des règles.

Dans la cadre de ce travail nous nous sommes intéressées par les méthodes basées sur l'apprentissage automatique, principalement ceux basé sur: les champs aléatoires conditionnels, modèle de Markov caché, modèle de Markov à entropie maximale, Les machines à vecteurs de support ou séparateurs à vaste marge, naïve bayésienne, L'entropie maximale et Arbre de Décision.

3. Modèles d'Apprentissages Automatiques

3.1. Modèle de Markov caché - HMM

Le HMM [78] est un type génératif de modèle basé sur une séquence. Dans l'explication suivante des modèles de séquence, x correspond à la séquence de jetons d'entrée et y à la séquence d'étiquettes de sortie. Les modèles génératifs trouvent généralement la meilleure séquence de points en calculant la forme générative de la probabilité p (x, y). En HMM, la probabilité de p (y | x) peut être réécrite sous forme de calcul utilisant sa forme générative p (x, y) conformément au théorème de Bayes :

$$p(y|x) = \frac{p(x,y)}{p(x)} \quad (2.1)$$

En supposant que l'étiquette actuelle yi dépend de la balise précédente yi-1, et que le jeton actuel xi dépend de l'étiquette actuelle yi, alors p (x, y) peut être réécrit comme suit :

$$p(x,y) = \prod_{i=1}^{n} p(y_{i-1}|y_i) * \prod_{i=1}^{n} p(x_i|y_i) \quad (2.2)$$

Où n est le nombre de jetons dans x.

Puisque l'objectif est de trouver le meilleur p (y | x), et que p (x) est une probabilité a priori qui reste la même pour chaque classe d'étiquettes possible, il suffit de comparer p (x, y).

Le problème de l'équation (2.2) réside dans la fragmentation des données causée par p (xi | yi). Idéalement, on aura suffisamment de données d'apprentissage pour chaque valeur possible de xi afin de calculer p (xi | yi). En réalité, il existe rarement assez de données d'apprentissage pour calculer des probabilités précises lors du décodage de nouvelles données. Ce problème est souvent résolu en utilisant l'approche Naïve bayésienne Dans l'approche Naïve bayésienne, nous décomposons p (xi | yi) comme suit :

$$p(x_i|y_i) = \prod_j p(f_{ij}|y_i) \quad (2.3)$$

Où f_{ij} est la valeur de la $j^{ème}$ caractéristique de x_i.

Même avec la solution ci-dessus, les HMM souffrent de deux limitations*. La première découle de l'hypothèse Naïve bayésienne de HMM pour résoudre les NER, qui bénéficieraient d'une

représentation plus riche des observations, en particulier d'une représentation décrivant les observations en termes de nombreuses caractéristiques superflues, telles que les majuscules, les affixes, l'étiquetage morpho-syntaxique (***Parts of Speech – POS***), en plus des caractéristiques de mot de surface.

3.2. Modèle de Markov à Entropie Maximale – MEMM

Les MEMM [76] ont été proposés pour aborder les deux problèmes susmentionnés (*). Pour permettre l'observation de caractéristiques non indépendantes difficiles à énumérer, MEMM remplace la paramétrisation générative de probabilité conjointe utilisée par les HMM par un modèle conditionnel qui représente la probabilité d'atteindre un état, basé sur une caractéristique d'observation et l'état précédent. La probabilité de p (y|x) dans le MEMM est calculée comme suit :

$$p(y|x) = \prod_i p(y_i|y_{i-1}, x) \tag{2.4}$$

$$p(y_i|y_{i-1}, x) = \frac{1}{Z(y_{i-1}, x)} exp \sum_j \lambda_j h_j(y_{i-1}, y_i, x, i) \tag{2.5}$$

$$Z(y_{i-1}, x) = \sum_{y_i} exp \sum_j \lambda_j h_j(y_{i-1}, y_i, x, i) \tag{2.6}$$

Où $Z(y_{i-1}, x)$ est le facteur de normalisation qui assure la probabilité de tout état y_i à un, $h_j(y_{i-1}, y_i, x, c)$ est généralement une fonction caractéristique binaire et λ_j est son poids. Les grandes valeurs positives λ_j indiquent une préférence pour un tel événement, alors que les grandes valeurs négatives font le cas peu probable.

3.3. Champs Aléatoires Conditionnels – CRF

Les CRF sont des modèles graphiques non dirigés, dans lesquels chaque nœud représente un état formé pour maximiser une probabilité conditionnelle [79]. Un CRF à chaîne linéaire avec des paramètres $\Lambda = \{\lambda 1, \lambda 2 \ldots\}$ définit une probabilité conditionnelle pour une séquence d'états $y=y_1 \ldots y_n$ à partir d'une longueur n séquences d'entrée $x=x_1 \ldots x_n$ comme suit:

$$p(y|x) = \frac{1}{Z(x)} exp \sum_{c \in C} \sum_j \lambda_j h_j(y_c, x, c) \tag{2.7}$$

$$Z(x) \equiv \sum_y exp \sum_{c \in C} \sum_j \lambda_j h_j(y_c, x, c) \tag{2.8}$$

Où Z(x) est le facteur de normalisation garantissant la somme de toutes les séquences d'états à un, ***C*** : est l'ensemble des cliques de la phrase cible, et ***c*** : est tout clique unique. Une clique est un sous-ensemble de nœuds entièrement connectés.

Notez que $h_j(y_c, x, c)$ est généralement une fonction caractéristique binaire et que λ_j est son poids. Des valeurs λ_j positives élevées indiquent une préférence pour un tel événement, tandis que des valeurs négatives importantes rendent l'événement improbable.

La taille de c détermine les états auxquels h_j peut se référer. Si c ne contient que l'état actuel, h_j ne peut se référer qu'à l'état actuel. Cependant, si c contient les états actuel et précédent, h_j peut se référer à tous. Étant donné x, la probabilité conditionnelle de y est égale à la somme exponentielle de $\lambda_j h_j$ dans toutes les cliques. Comme la longueur moyenne d'une NE biomédicale est comprise entre deux et trois jetons, nous en utilisons deux au lieu de trois pour la taille de la clique afin d'économiser l'espace mémoire. Voici deux exemples de fonctionnalités de mot actuelles. Le premier se réfère au couple y_i et y_{i-1} et le second au yi individuellement :

$$h_1(y_c, x, [i-1, i]) = \begin{cases} 1\,, & if y_{i-1} = B - protein and y_i = I - protein \\ & and Current\_word(x_i) = "protein" \\ 0, & otherwise \end{cases} \quad (2.9)$$

$$h_2(y_c, x, [i-1, i]) = \begin{cases} 1\,, & if y_i = I - protein \\ and Current\_word(x_i) = "protein" & \\ 0\,, & otherwise \end{cases} \quad (2.10)$$

Où y_c représente la séquence de balises dans c, et [i-1, i] signifie que la clique c va de i-1ème jeton au ième jeton.

3.4. Séparateurs à Vaste Marge – SVM

Comme d'autres méthodes de la même approche d'apprentissage automatique, SVM [80] prend en entrée un ensemble d'exemples d'apprentissage (donnés sous forme de vecteurs d'entités à valeur binaire) et cherche une fonction de classification qui les projettes à une classe.

Plusieurs points concernant les modèles SVM méritent d'être résumés ici. La première est que les SVM sont connus pour gérer de manière robuste de grands ensembles de fonctionnalités et pour développer des modèles qui maximisent leur généralisabilité. Cela en fait un modèle idéal pour la tâche BioNER. La généralisabilité dans les SVM repose sur la théorie de l'apprentissage statistique et sur l'observation, selon laquelle il est parfois utile de classer de manière erronée certaine données d'apprentissage afin de maximiser la marge entre les autres points d'apprentissage [81]. Ceci est particulièrement utile pour les ensembles de données du monde réel qui contiennent souvent des points de données inséparables. Deuxièmement, bien que l'apprentissage soit généralement lent, le modèle résultant est généralement petit et fonctionne rapidement car seuls les vecteurs de support doivent être conservés, c'est-à-dire les modèles qui aident à définir la fonction qui sépare les exemples positifs des exemples négatifs. Troisièmement, les SVM sont des classificateurs binaires et nous devons donc combiner des modèles SVM pour obtenir un classificateur multiclass.

Formellement, nous pouvons alors considérer que le but du SVM est d'estimer une fonction de classification f: $\chi \rightarrow \{\pm 1\}$ à l'aide d'exemples d'apprentissage tirés de $\chi \times \{\pm 1\}$ de manière à

minimiser l'erreur sur les exemples non vus. La fonction de classification renvoie soit +1 si les données de test sont membres de la classe, soit -1 si ce n'est pas le cas.

Bien que les SVM apprennent ce qui est essentiellement des fonctions de décision linéaires, l'efficacité de la stratégie est assurée en mappant les modèles d'entrée χ sur un espace de fonctions Γ à l'aide d'une fonction de mappage non linéaire : $\Phi : \chi \rightarrow \Gamma$.

3.5. Entropie Maximale – ME

Le Framework d'entropie maximal [82] estime les probabilités selon le principe de formuler le moins possible d'hypothèses hormis les contraintes imposées. Ces contraintes sont dérivées des données d'apprentissage et expriment une relation entre les caractéristiques et les résultats. La distribution de probabilité qui satisfait le principe ci-dessus est celle avec l'entropie la plus élevée. Il est unique, en accord avec la distribution du maximum de vraisemblance et sous la forme exponentielle (Della et *al.*) [83]:

$$p(o|h) = \frac{1}{Z(h)} \prod_{j=1}^{k} \alpha_j^{f_j(h,o)} \tag{2.11}$$

Où o indique le résultat, h l'historique (ou le contexte) et $Z(h)$ est une fonction de normalisation. De plus, chaque fonction caractéristique $f_j(h, o)$ est une fonction binaire. Par exemple, pour prédire si un mot appartient à une classe de mots, o est vrai ou faux, et *h* indique le contexte entourant :

$$f_j(h, o) = \begin{cases} 1 & if\ o = true, previousword = the \\ 0 & othwewise \end{cases} \tag{2.12}$$

Les paramètres α_j sont estimés par une procédure appelée échelle généralisée itérative (GIS) (Darroch et Ratcliff) [84]. C'est une méthode itérative qui améliore l'estimation des paramètres à chaque itération.

3.6. Arbre de Décision – DT

Les algorithmes d'arbre de décision [85] commencent par un ensemble d'observations ou d'exemples et créent une structure de données arborescente pouvant être utilisée pour classifier de nouvelles observations. Chaque cas est décrit par un ensemble d'attributs (ou caractéristiques) pouvant avoir des valeurs numériques ou symboliques.

Chaque cas d'apprentissage est associé à une étiquette représentant le nom d'une classe. Chaque nœud interne d'un arbre de décision contient un test dont le résultat est utilisé pour décider quelle branche à suivre de ce nœud. Par exemple, un test peut demander "x> 4 pour l'attribut x?" Si le test est vrai, le cas se poursuivra dans la branche gauche, sinon, il suivra vers la branche

droite. Les nœuds feuilles contiennent des étiquettes de classe au lieu de tests. En mode de classification, lorsqu'un scénario de test (sans étiquette) atteint un nœud feuille.

3.7. Classification Naïve Bayésienne – NB

NB est un classificateur probabiliste simple basé sur l'application du théorème de Bayes et sa méthode puissante, simple et indépendante du langage. Lorsque le classificateur NB est appliqué au problème de Catégorisation, nous utilisons l'équation (2.13).

$$p(class|document) = \frac{p(class).p(document|class)}{p(document)} \quad (2.13)$$

Où p(class|document) C'est la probabilité de classe d'un document, ou la probabilité qu'un document D appartient à une classe C.

p(document): La probabilité d'un document, nous pouvons remarquer que p(document) est un diviseur de constance pour chaque calcul, nous pouvons donc l'ignorer.

p(class): Probabilité d'une classe (ou catégorie), on peut la calculer à partir du nombre de documents de la catégorie divisé par le nombre de documents dans toutes les catégories.

p(document|class): représente la probabilité d'un document de classe donnée et les documents peuvent être modélisés comme des ensembles de mots, ainsi le p(document|class) peut-être écrit comme suit :

$$p(document|class) = \prod_i p(word_i|class) \quad (2.14)$$

Donc :

$$p(document|class) = p(class).\prod_i p(word_i|class) \quad (2.15)$$

Où : $p(word_i|class)$: La probabilité que le i-ème mot d'un document donné apparaisse dans un document de la classe C, peut être calculée comme suit:

$$p(word_i|class) = \frac{Tct+\alpha}{Nc+V} \quad (2.16)$$

Où : Tct est le nombre de fois que le mot apparaît dans cette catégorie ; Nc est Le nombre de mots dans la catégorie ; V est la taille du tableau de vocabulaire et α est la constante positive, généralement 1 ou 0,5, pour éviter une probabilité nulle.

3.8. Avantages et faiblesses des méthodes d'apprentissage automatique

Dans cette section, nous présentons une étude théorique comparative entre les sept modèles ML décrits ci-dessus.

Les CRF présentent plusieurs avantages par rapport aux autres méthodes. Premièrement, son principal avantage par rapport à HMM est sa nature conditionnelle, ce qui permet d'assouplir les hypothèses d'indépendance requises par HMM pour garantir des inférences raisonnables. De plus,

le CRF évite le problème de biais d'étiquette [79] inhérent à MEMM [86] et à d'autres modèles de Markov conditionnels basés sur des modèles graphiques dirigés [87]. Le CRF surpasse à la fois les MEMM et les HMM dans un certain nombre de tâches d'étiquetage de séquence dans le monde réel [79] [88] [89]. En outre, CRF utilise des sommes pondérées de manière exponentielle pour combiner les influences de nombreuses entités corrélées, y compris des entités superposées et co-dépendantes. En conséquence, nous pouvons utiliser plusieurs fonctionnalités avec CRF plus facilement qu'avec HMM. La différence critique entre CRF et MEMM réside dans le fait que ce dernier utilise des modèles exponentiels par état pour les probabilités conditionnelles des états suivants compte tenu de l'état actuel, alors que CRF utilise un modèle exponentiel unique pour déterminer la probabilité conjointe de la séquence complète des étiquettes, compte tenu de la séquence d'observation. Par conséquent, dans CRF, les poids de différentes entités dans différents états rivalisé les uns contre les autres.

Le contexte de caractéristique utilisé par SVM est global et il n'a pas les mêmes contraintes que les CRFs. SVM est initialement la meilleure solution pour les tâches de classe binaire et ne fonctionne pas correctement pour les tâches de classe multiple. Les CRFs nécessitent généralement plus de temps et d'espace de calcul que les SVM. Ainsi, bien que les CRF présentent des inconvénients, ils sont très efficaces pour les tâches d'étiquetages de données séquentielles, ce qui est un problème typique du NER [90].

En outre, CRF contourne les limitations de ME, DT et NB. La vraisemblance maximale exacte de La solution d'entropie maximale n'existe pas dans certaines circonstances. Pour les données comprenant des variables catégorielles avec un nombre différent de niveaux, Les informations obtenues dans les arbres de décision sont biaisées en faveur de ces attributs avec plus de niveaux. Enfin, CRF peut utiliser plusieurs fonctionnalités plus facilement que NB et les autres modèles.

Enfin, les avantages et les faiblesses des sept méthodes présentées dans cette section peuvent être résumés comme suit :

Tableau 1: Avantages et faiblesses des méthodes d'apprentissage automatique

Méthodes D'apprentissage Automatique	**Avantages**	**Faiblesses**
Champs Aléatoires Conditionnels (CRFs)	✓ Est une approche plutôt moderne qui est déjà devenue très populaire pour un grand nombre de tâches de la NLP ✓ Evite le problème de biais d'étiquette ✓ Peut utiliser plusieurs fonctionnalités plus facilement que d'autres.	× Utilise une taille de contexte limitée plutôt que le texte entier. × Affecté par la distribution des données.
Modèle de Markov Caché (HMM)	✓ Offre un apprentissage rapide ✓ Offre une maximisation globale de la probabilité conjointe sur l'ensemble des séquences d'observation et d'étiquetage.	× Assume que chaque mot soit indépendant de son contexte. × Il définit ses paramètres pour maximiser la vraisemblance de la séquence d'observation.
Modèle de Markov à Entropie Maximale (MEMM)	✓ Permettre observation non indépendants des caractéristiques qui sont difficiles à énumérer.	× Les hypothèses markoviennes dans les MEMMs ignorent les dépendances entre l'état actuel et les autres états.
Séparateurs à Vaste Marge (SVM)	✓ Adéquate aux problèmes de classification binaire. ✓ Ne nécessite pas grand domaine	× Plus de temps nécessaire pour entrainer un modèle complexe. × Ne fonctionne pas bien sur des tâches à plusieurs classes.
Classification Naïve Bayésienne (NB)	✓ Rapide à entrainer (scan unique). Rapide à classer. ✓ Non sensible aux caractéristiques non pertinentes.	× Assume l'indépendance des caractéristiques.
L'Entropie Maximale (ME)	✓ Il permet d'utiliser des caractéristiques plus flexibles que naïve Bayésienne et elle est capable d'utiliser plus de preuves pour chaque prédiction que la technique des arbres de décision.	× La principale limitation est que la vraisemblance maximale exacte de La solution d'entropie maximale n'existe pas dans certaines circonstances
Arbre de Décision (DT)	✓ Avoir la valeur même avec peu de données. D'importantes leçons peuvent être générées à partir d'experts décrivant une situation et leurs préférences pour de résultats.	× Pour les données comprenant des variables catégorielles avec un nombre différent de niveaux, Les informations obtenues dans les arbres de décision sont biaisées en faveur de ces attributs avec plus de niveaux

4. Organigramme du système d'évaluation

La figure 1 présente l'organigramme général d'un système Bio-NER utilisant une approche d'apprentissage automatique.

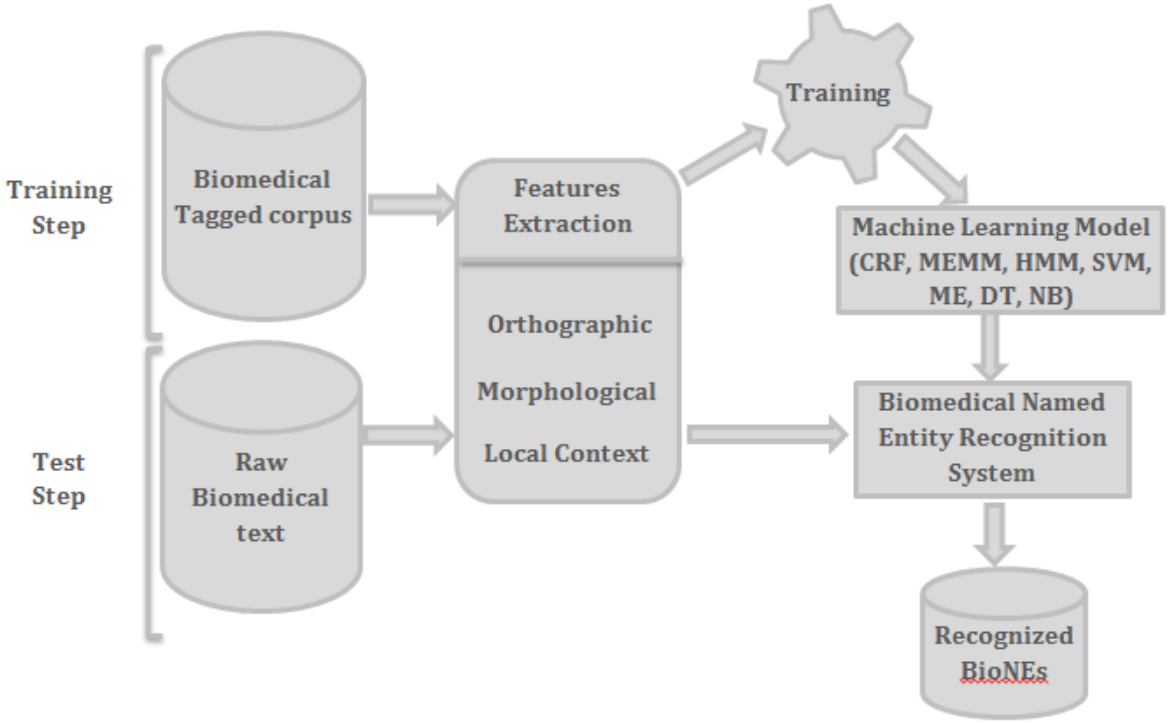


Fig 1 Organigramme du système BioNER basé sur un modèle d'apprentissage automatique

En plus, pour rendre les applications d'apprentissage automatique plus sophistiquées, nous incluons des routines de transformation des documents texte en représentations numériques pouvant ensuite être traitées efficacement. Ce processus est mis en œuvre via un système flexible de "Pipe", qui gère des tâches distinctes telles que la création de chaînes symboliques, la suppression de mots vides et la conversion de séquences en vecteurs de comptage.

Pour entraîner des modèles d'apprentissage automatique, toutes les phrases doivent être divisées en jetons, chaque jeton dans les données d'apprentissage devant être identifié comme faisant partie ou non du nom d'une entité nommée biomédical (c'est-à-dire donnant des tags: NE (entité nommée) ou NNE (NON Entité nommée)

5. Expérimentations et Résultats

Dans cette section, nous présentons expérimentalement une étude comparative des sept méthodes d'apprentissage automatique. Nous introduisons d'abord les métriques de performance, les données d'expérimentation, la configuration expérimentale, puis communiquons les résultats d'évaluation détaillés de notre étude.

5.1. Corpus

Plusieurs corpus disponibles peuvent être utilisés pour entrainer et évaluer les systèmes de NER. Pour permettre une comparaison directe entre les différentes méthodes d'apprentissage automatique (CRF, HMM, MEMM, SVM, ME, NB, DT) utilisées dans cette étude, nous utilisons deux des corpus les plus utilisés : GENIA [91] et JNLPBA [92]. Le corpus GENIA a été formé par une recherche contrôlée de MEDLINE à l'aide des termes MeSH «humain», «cellules sanguines» et «facteurs de transcription», 2000 résumés ont été sélectionnés et annotés manuellement. D'autre part, le corpus JNLPBA contient 2404 résumés extraits de MEDLINE. L'annotation manuelle de ces résumés repose sur cinq classes de l'ontologie GENIA [93], à savoir les protéines, l'ADN, l'ARN, la lignée cellulaire et le type de cellule.

5.2. Évaluation des performances

Pour analyser les résultats obtenus par les techniques d'apprentissage automatique étudiées auparavant et comparer leurs performances, nous utilisons des métriques d'évaluation communes :

- Précision (c'est-à-dire valeur prédictive positive), capacité d'un système à ne présenter que des éléments pertinents ;

$$P = \frac{TP}{TP+FP} \tag{2.17}$$

- Rappel (c'est-à-dire, la sensibilité) la capacité d'un système de présenter tous les éléments pertinents ;

$$R = \frac{TP}{TP+FN} \tag{2.18}$$

- F-mesure, la moyenne harmonique de précision et de rappel. Ces mesures sont formulées comme suit :

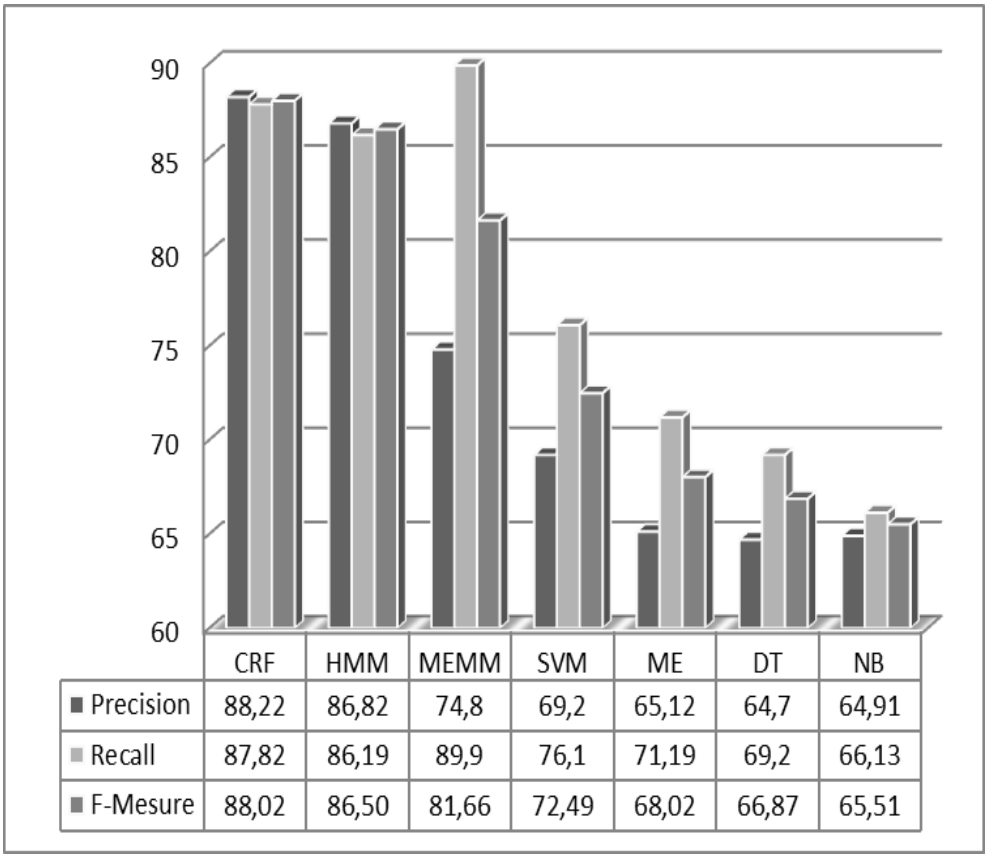
$$F1\ measure = 2\frac{P*R}{P+R} \tag{2.19}$$

Où : True Positive (TP) fait référence aux BioNE détectées correctement en tant que BioNE. Faux Positif (FP) fait référence aux BioNE détectés et qui ne sont pas des BioNE. Faux négatif (FN) fait référence aux BioNEs qui ne sont pas détectés.

5.3. Résultats

La figure 2 présente les résultats obtenus sur le corpus GENIA, en comparant les modèles CRF, HMM, MEMM, SVM, ME, DT et NB. La CRF surpasse toutes les autres méthodes d'apprentissage automatique avec une valeur de F-Measure égale à 88,02%. Elle présente une amélioration significative de 1,52% par rapport à la deuxième meilleure méthode HMM. En comparaison

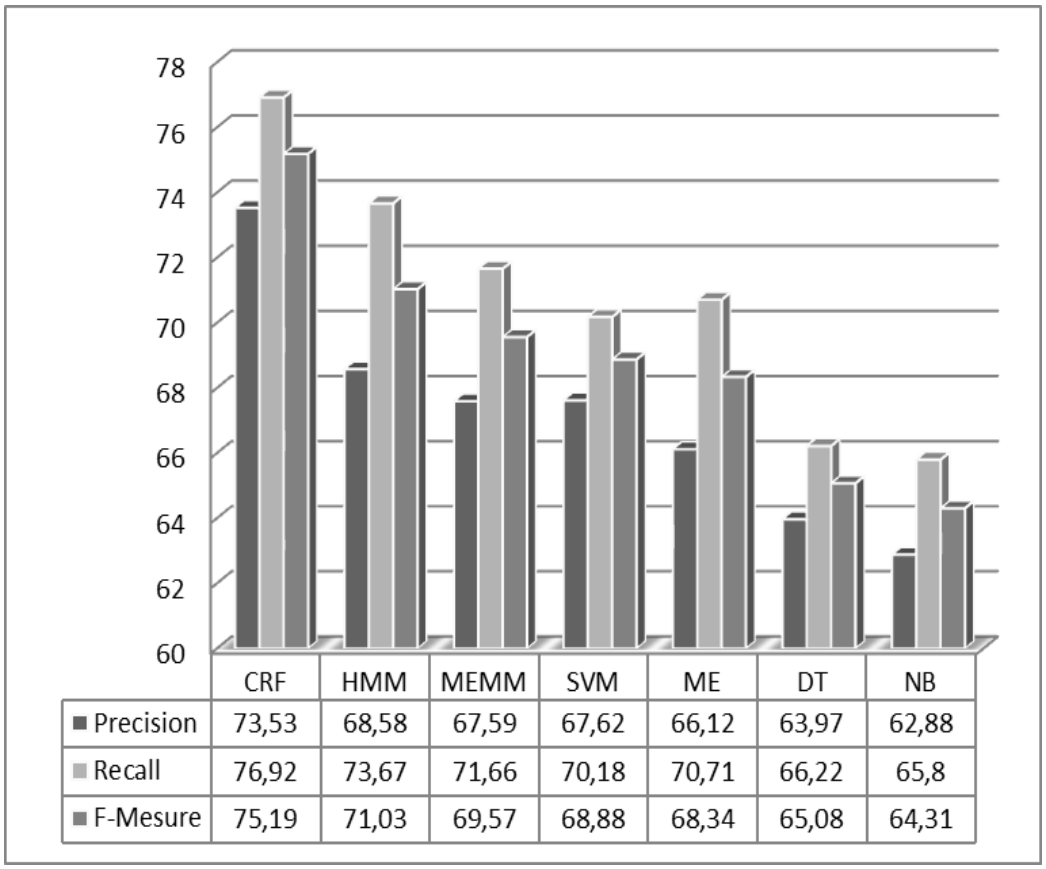
avec la troisième meilleure méthode (MEMM), la CRF présente une amélioration de 6,36%. Comparé à la méthode la moins efficace (NB), la CRF présente une amélioration de 22,51%. Globalement, la CRF présente les meilleurs résultats à la fois en termes de précision et de rappel.



	CRF	HMM	MEMM	SVM	ME	DT	NB
Precision	88,22	86,82	74,8	69,2	65,12	64,7	64,91
Recall	87,82	86,19	89,9	76,1	71,19	69,2	66,13
F-Mesure	88,02	86,50	81,66	72,49	68,02	66,87	65,51

Fig 2 Comparaison des Performances dans le corpus GENIA

De même, la figure 3 présente les résultats obtenus sur le corpus JNLPBA. Aussi, La CRF surpasse toutes les autres méthodes d'apprentissage automatique, avec une valeur de F-Measure égale à 75,19%, elle présente une amélioration significative de 4,16% par rapport à la deuxième meilleure méthode HMM. En comparaison avec la troisième meilleure méthode MEMM, la CRF présente une amélioration de 5,62%. Comparé à la méthode la moins efficace NB, la CRF présente une amélioration de 10,88%. Globalement, elle présente les meilleurs résultats à la fois en précision et en rappel.



	CRF	HMM	MEMM	SVM	ME	DT	NB
Precision	73,53	68,58	67,59	67,62	66,12	63,97	62,88
Recall	76,92	73,67	71,66	70,18	70,71	66,22	65,8
F-Mesure	75,19	71,03	69,57	68,88	68,34	65,08	64,31

Fig 3 Comparaison des Performances dans le corpus JNLPBA

Globalement, le modèle CRF surpasse de manière significative toutes les autres méthodes d'apprentissage automatique sur GENIA et JNLPBA, en adaptant simplement la liste des caractéristiques utilisé pour chaque corpus et type d'entité.

Le tableau 2 présente des exemples Bio-NE reconnues par le modèle CRF, mais ne sont pas reconnus par HMM et MEMM. Ces exemples sont une preuve de la discussion à la sous-section 3 sur les avantages du CRF et les raisons pour lesquelles il offre de meilleures performances que d'autres modèles.

Dans le tableau 3, nous présentons également d'autres exemples de BioNE reconnues par le modèle HMM, mais ne sont pas reconnu par MEMM.

Tableau 2: Comparaison de Bio-NER entre les méthodes CRF, HMM et MEMM en utilisant GENIA corpus

	Modèle		
	CRF	***HMM***	***MEMM***
Intact promoter	OUI	NON	NON
Monoclonal antibodies	OUI	NON	NON
Transfection assays	OUI	NON	NON
Therapeutic strategies	OUI	NON	NON
Adenovirus (Ad) infection	OUI	NON	NON

Tableau 3 Comparaison de Bio-NER entre les méthodes HMM et MEMM en utilisant GENIA corpus

	Modèle	
	HMM	***MEMM***
Lipoxygenase	OUI	NON
CD28 surface receptor	OUI	NON
Cell proliferation	OUI	NON
Purine-rich binding sites	OUI	NON
PuB1	OUI	NON

6. Conclusion

Dans ce chapitre nous avons étudié dans un premier temps le système de reconnaissance des entités nommées Biomédicale, dont nous avons présenté une étude théorique et expérimentale comparative entre sept méthodes d'apprentissage automatique pour Bio-NER (modèles CRF, HMM, MEMM, SVM, ME, NB et DT). Nous avons résumé leurs avantages et leurs faiblesses et expliqué pourquoi les CRF présentent plusieurs avantages par rapport à d'autres méthodes. Afin d'évaluer ces techniques et de les comparer, nous avons utilisé deux corpus bien connus : GENIA et JNLPBA. Les expériences ont prouvé que le modèle CRF avait obtenu les meilleurs résultats de mesure de F1 de 88,02% et 75,19% sur chaque corpus, respectivement.

Avec les résultats obtenus et les caractéristiques présentées par notre étude, nous concluons que le CRF est une technique d'apprentissage automatique de pointe pour la reconnaissance des entités nommées biomédicale, contribuant à une représentation de document par Sac de Mot de meilleure qualité.

Chapitre 3 : Résolution de l'Ambigüité dans la Représentation Conceptuelle

1. Introduction

La manière habituelle de représenter un texte est un ***Sac de mots – BoW***. Cette représentation souffre du manque de sens dans la représentation résultante qui ignore toute la sémantique qui réside dans le texte original ; L'intégration de la sémantique dans la représentation textuelle est résolue par l'utilisation des concepts - en tant qu'unité de connaissance. Nous nous référons à cette intégration par un processus de conceptualisation, ce dernier consiste à effectuer une étape de mappage des termes de documents aux concepts correspondants dans la ressource sémantique. Lors de la recherche dans la ressource sémantique pour le mappage d'un mot polysémique, la conceptualisation peut trouver plusieurs correspondances ayant des significations différentes. Par exemple, le mot "Book" signifie en anglais un livre et une réservation (Ticket, hébergement, etc.). Pour faire face à ce problème, la Résolution de l'Ambiguïté Lexicale – WSD identifie automatiquement le sens approprié d'un mot ambigu en fonction du contexte dans lequel le mot est utilisé.

Pour la désambiguïsation des sens basée sur la connaissance non supervisée, afin de déterminer le sens correct (concept correct), on utilise généralement des mesures de similarité sémantique ou de connexité [65] qui tentent de quantifier la proximité sémantique entre deux concepts.

Dans le cadre de ce chapitre, nous étudions dans un premier temps l'intégration de la sémantique dans la représentation par la conceptualisation. Nous avons choisi l'utilisation d'une première étape de reconnaissances des entités nommées biomédicales avant d'entamer le processus de conceptualisation afin de tirer parti des résultats issues de notre étude concernant la BioNER présenté dans le chapitre précédent, ainsi dans l'approche proposée, nous appliquons la conceptualisation aux termes issues de système de BioNER à travers l'enrichissement par les concepts de l'ontologie utilisée. Nous allons également discuter les différentes stratégies de conceptualisation et de désambiguïsation et présenter un cadre de travail générique pour la conceptualisation du texte biomédical. Ensuite nous présentons une étude concernant les mesures de similarité entre deux concepts, cette dernière est une évaluation d'une méthode utilisée pour déterminer le concept de mot ambigu le plus approprié au contexte. Nous évaluons également l'effet de la taille de la fenêtre de contexte sur cette méthode, en utilisant le degré de similarité sémantique et de connexité entre les concepts possibles et les termes entourant le mot ambigu.

2. Conceptualisation du Texte Biomédical

Dans l'objectif de surmonter les limites de la représentation basée sur les mots, celle basée sur concepts utilise des ressources sémantiques, telles que des thésaurus et des ontologies, pour remplacer la représentation basée sur les termes par une représentation basée sur un concept. Ainsi,

cette présentation basée sur concept permet d'enrichir sémantiquement la représentation finale du document.

La conceptualisation est par définition : "Interpréter de manière conceptuelle" [94]. Dans le contexte de l'analyse du texte, il s'agit du processus consistant à mapper des mots littéralement présents dans du texte avec leurs concepts ou leurs sens correspondants, en les faisant correspondre à des ressources sémantiques. L'application de l'indexation au texte conceptualisé pourrait améliorer les résultats de la classification [95].

Selon la littérature, la conceptualisation a été appliquée à des mots en utilisant différentes stratégies [96]. Comme exemple de ressources sémantiques utilisées pour la conceptualisation : WordNet [97], Wikipedia [98] et d'autres ressources spécifiques à un domaine, appelées habituellement ontologies de domaine, telles que UMLS [36], dans le domaine médical. En général, la conceptualisation du texte est réalisée en deux étapes :

- Analyser le texte afin de trouver des mots candidats pour la correspondance mot à concept.
- Rechercher les concepts correspondants liés aux mots candidats, puis intégrer ces concepts dans un texte produisant le texte conceptualisé final.

La sous-section suivante présente les différentes stratégies de conceptualisation ou les différentes manières d'intégrer les concepts mappés du texte conceptualisé final. Ensuite, nous présentons différentes stratégies pour faire face aux ambiguïtés.

2.1. Stratégies De Conceptualisation De Texte

Lors de la conceptualisation, nous mappons les mots de texte aux concepts correspondants dans la ressource sémantique. La prochaine étape consiste à incorporer ces concepts dans le texte résultant. Selon la littérature, trois stratégies différentes sont possibles pour conceptualiser les vecteurs de mots [99]. Nous adaptons ces stratégies à notre approche de la conceptualisation de texte comme suit :

- Ajout de concepts (***Adding Concepts***) : cette stratégie étend le texte d'origine en ajoutant les concepts mappés. Le texte conceptualisé contient les mots originaux ainsi que ces concepts [100].
- Conceptualisation partielle (***Partial Conceptualization***) : Cette stratégie substitue les mots par les concepts correspondants et conserve les mots n'ayant pas de concepts correspondant dans le texte. Le texte conceptualisé contient les concepts mappés et quelques mots originaux [95].
- Conceptualisation complète (***Complete Conceptualization***) : De la même manière que la conceptualisation partielle, cette stratégie substitue également les mots par les concepts. La

principale différence est qu'elle élimine les mots restants dans le texte conceptualisé final qui ne devrait contenir que des concepts [101].

Selon les auteurs [102], la deuxième stratégie semble être la plus appropriée, car elle remplace les mots par les concepts associés, de sorte qu'aucune caractéristique originale n'est supprimée du texte (par rapport à la troisième). Aucune fonctionnalité supplémentaire n'est ajoutée (par rapport à la première), ce qui minimise les effets sur l'efficacité. Cependant, ce n'est pas le choix des auteurs [95] qui ont utilisé Ajout de concepts ou les auteurs [101] qui ont utilisé la conceptualisation Complète.

2.2. Stratégies de Désambiguïsation

Lors de la recherche dans la ressource sémantique pour le mappage d'un mot polysémique, la conceptualisation peut trouver plusieurs correspondances ayant des significations différentes. Par exemple, le mot "Book" signifie en anglais un livre et une réservation (Ticket, hébergement, etc.). Pour faire face à ce problème, les approches de pointe en matière de conceptualisation proposent de multiples stratégies pour faire face aux ambiguïtés [99]. Nous citons ici trois stratégies de désambiguïsation qui peuvent aider à résoudre ce problème :

- Tous les Concepts (***All Concepts***) : Cette stratégie accepte tous les concepts candidats en tant que correspondance pour le mot considéré.
- Premier Concept (***First Concept***) : Cette deuxième stratégie accepte le concept le plus fréquemment utilisé parmi les différents candidats selon les statistiques linguistiques.
- Contexte (***Context***) : Cette troisième stratégie accepte les concepts candidats ayant le contexte sémantique le plus similaire par rapport au contexte du mot d'origine dans le document [101] [103].

La première stratégie, bien qu'elle soit la plus simple, est la moins fiable, car elle accepte tous les concepts candidats sans choisir le sens approprié du mot. La deuxième stratégie est plus fiable. Néanmoins, cette stratégie ne permet pas de choisir le bon concept si le sens correct correspond au sens rarement utilisé du mot. Malgré sa complexité, la dernière stratégie semble être la plus fiable et la plus précise [99] et elle a été utilisé par la plupart des approches de pointe qui traitent le problème de l’ambiguïté [101] [103] .Le contexte d'un concept est lié à sa définition ou à ses mots descriptifs dans la ressource sémantique ou à son contexte textuel dans un corpus de texte.

3. Cadre générique pour la conceptualisation du texte

Dans la section précédente nous avons présenté différentes stratégies de conceptualisation et de résolution de l'ambiguïté lors de la conceptualisation. Cette section présente un cadre générique pour la conceptualisation de texte à travers différentes étapes de traitement (Figure 4).

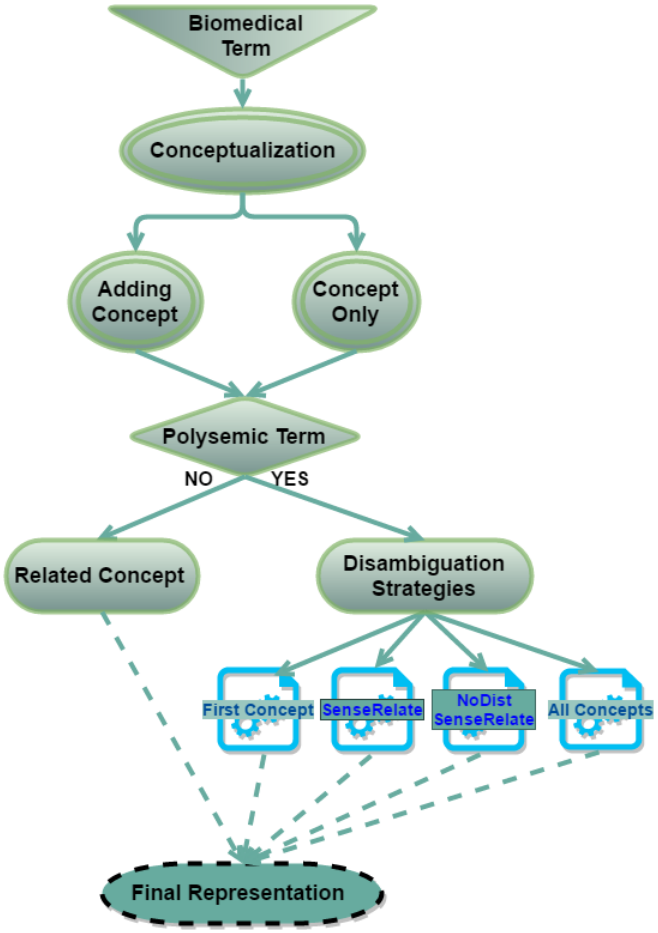


Fig 4 Plate-forme générique pour la conceptualisation de texte

La première étape décompose le texte en jetons et identifie les N-grammes candidats à l'appariement des concepts. Dans cette étape, nous utilisons la reconnaissance des entités nommées décrite dans la section 2. Ensuite, l'étape de recherche des correspondances dans la ressource sémantique pour chacun des candidats. Ces correspondances sont lexicalement similaires aux candidats ou à leurs dérivations. Si le système trouve plusieurs correspondances pour le même candidat, il applique une stratégie de désambiguïsation pour résoudre cette ambiguïté afin de choisir la correspondance correcte. Enfin, le système intègre les concepts correspondants dans le texte d'origine conformément à la stratégie de conceptualisation afin de produire le texte conceptualisé final. Nous avons choisi l'utilisation d'une première étape de reconnaissances des entités nommées biomédicales avant d'entamer le processus de conceptualisation afin de tirer parti des résultats issues de notre étude concernant la BioNER présentée dans la section précédente. Ce cadre est générique

et modulaire ; différentes techniques et différents domaines d'application peuvent s'intégrer dans le système.

Dans la suite du chapitre, nous présentons en détails la désambiguïsation du sens de mot, ses approches ainsi les mesures de similarité et de connexité sémantique utilisées pour résoudre l'ambiguïté.

4. Désambiguïsation du Sens de Mot

La désambiguïsation du sens des mots - WSD est un problème ouvert du traitement du langage naturel (***Natural Language Processing – NLP***) qui tente d'identifier le sens correct d'un mot afin de résoudre les ambiguïtés lexicales. Elle consiste à identifier automatiquement le sens approprié d'un mot ambigu en fonction du contexte dans lequel le mot est utilisé. Par exemple, le terme "cold" (froid) peut faire référence à la température ou au rhume, en fonction de la manière dont le mot est utilisé dans la phrase. L'identification automatique du sens voulu des mots ambigus améliore les performances des applications biomédicales et cliniques telles que les applications de codage et d'indexation médicales qui deviennent des tâches essentielles en raison de la quantité croissante d'informations disponibles pour les chercheurs.

WSD joue un rôle essentiel dans de nombreuses applications sur le texte biomédical telles que l'extraction d'informations, la catégorisation de texte et la traduction automatique. Pour le domaine biomédical, la Bibliothèque Nationale de Médecine (***National Library of Medicine – NLM***) a conclu que l'ambiguïté lexicale dans le Système de Langage Médical Unifié (***Unified Medical Language System – UMLS***) était le principal défi de l'indexation des revues biomédicales avec des concepts du métathésaurus UMLS [36].

Historiquement, il a été démontré que les approches d'apprentissage automatique supervisé désambiguïsaient les termes avec un degré de précision plus élevé que les méthodes non supervisées. L'inconvénient des méthodes supervisées est qu'elles requièrent des données d'apprentissage annotées manuellement pour chaque terme devant faire l'objet d'une ambiguïté. Toutefois, l'annotation manuelle est un processus coûteux, difficile et long qui n'est pas pratique à appliquer à grande échelle. De ce fait, les méthodes non supervisées sont plus appréciées par les chercheurs du domaine, vue qu'elles sont non couteuses en terme d'annotation et qui peuvent se baser sur des Ressources de Connaissance Externe du domaine médical.

Pour la désambiguïsation des sens basée sur la connaissance non supervisée, afin de déterminer le sens correct (concept correct), nous utilisons généralement des mesures de similarité sémantique ou de connexité qui tentent de quantifier la proximité sémantique entre deux concepts.

Des études antérieures ont comparé et ont évalué l'efficacité de la mesure de la similarité sémantique et de la connexité sur le WSD biomédical [65].

Dans ce chapitre, nous réalisons une étude évaluative d'une méthode utilisée pour déterminer le concept de mot ambigu le plus approprié au contexte. Nous évaluons également l'effet de la taille de la fenêtre de contexte sur cette méthode, en utilisant le degré de similarité sémantique et de connexité entre les concepts possibles et les termes de son contexte, c'est à dire les termes qui entourent le mot ambigu.

5. Approches de la désambiguïsation de sens du terme

Il existe de nombreuses approches proposées qui tentent de résoudre le problème de la WSD. Les méthodes existantes peuvent être classées en trois groupes : les méthodes supervisées [104], non supervisées [105] et basées sur la connaissance [106].

Afin de déterminer le concept adéquat du mot ambigu pour les méthodes supervisées, nous utilisons des algorithmes d'apprentissage automatique. Bien évidement dans l'apprentissage automatique, les données d'apprentissage doivent être créées manuellement ou automatiquement pour chaque mot ambigu, ce qui est irréalisable à grande échelle.

Les méthodes non supervisées n'utilisent pas de données d'apprentissage, mais utilisent les caractéristiques de distribution d'un corpus externe. Pour les méthodes basées sur la connaissance, nous utilisons des ressources sémantiques externes et éventuellement des informations depuis corpus.

Généralement, les approches d'apprentissage supervisées surpassent les approches non supervisées [107], mais dans le domaine biomédical, il est très coûteux de créer un corpus annoté manuellement à des fins d'algorithme, ce qui rend l'approche non supervisée plus pratique [108]. Dans ce travail.

5.1. Méthodes basées sur Ressource de Connaissance Externe

Alexopoulou et *al.* [109] ont introduit la méthode dite du «sens le plus proche (Closest Sense)» qui tente de calculer la distance moyenne la plus courte entre le type sémantique d'un concept possible et les types sémantiques de chacun des mots entourant le mot cible. Ceci, est fait pour chaque concept possible et le concept avec la distance la plus courte est attribué au mot cible.

Jimeno-Yepes et *al.* [110] ont présenté également la méthode AEC, dans laquelle les instances de mot cible ont été formées pour générer automatiquement des données d'apprentissage à partir de MEDLINE. MEDLINE est indexé manuellement avec les termes MeSH, chaque terme étant associé à une CUI (***Concept Unique Identifier***) dans le système UML. Les citations de

MEDLINE contenant le mot cible et annotées avec l'un des sens possibles du mot cible sont extraites. Ces citations sont utilisées comme données d'apprentissage dans un algorithme WSD supervisé.

Garla et Brandt [111] ont utilisé l'algorithme SenseRelate proposé par Patwardhan et *al.* [112] pour évaluer les mesures de similarité fondées sur le chemin et la taxonomie. Dans cette méthode, on attribue à chaque concept possible de mot ambigu un score en faisant la somme du score de similarité entre celui-ci et les termes qui l'entourent.

6. Mesures de similarité sémantique et de connexité

Les mesures de connexité quantifient le degré d'association de deux mots (papier-ciseaux). La similarité est un sous-ensemble de relations et quantifie la similitude de deux concepts en fonction de leur emplacement dans une hiérarchie est-un (is-a) (voiture-véhicule). Cette section décrit les mesures de similarité et de connexité utilisées dans ce travail.

6.1. Mesures de similarités

Les mesures de similarité sémantique existantes peuvent être classées en deux groupes : basées sur trajectoire et basées sur le Contenu Informatif (***Information Content – IC***). Les mesures basées sur la trajectoire reposent sur les informations de chemin le plus court et les mesures basées sur IC incorporant la probabilité que le concept se produise dans un corpus de texte.

A. Basée sur trajectoire (Path-based)

Rada et *al.* [113] ont introduit la mesure de distance conceptuelle qui est la longueur du plus court chemin entre deux concepts (c1 et c2) dans MeSH en utilisant des relations RB / RN. Caviedes et Cimino [114] ont ensuite évalué cette mesure en utilisant les relations PAR / CHD. La mesure de trajectoire est une modification de celle-ci et est calculée comme l'inverse de la longueur du chemin le plus court.

Wu et Palmer [115] ont étendu cette mesure en incorporant la profondeur du plus petit subsume inférieur (LCS). Le LCS est le concept le plus spécifique que deux concepts partagent en tant qu'ancêtre. Dans cette mesure, la similarité est le double de la profondeur des deux concepts LCS divisée par le produit des profondeurs des concepts individuels tels que définis dans l'équation. (3.1).

$$sim_{wup} = \frac{2*depth(lcs(c_1,c_2))}{depth(c_1)+depth(c_1)} \quad (3.1)$$

Leacock et Chodorow [116] ont étendu la mesure du trajet en incorporant la profondeur de la taxonomie. Ici, la similarité est le log négatif du chemin le plus court (minpath) entre deux concepts divisé par le double de la profondeur totale de la taxonomie (D) telle que définie dans l'équation. (3.2).

$$sim_{lch} = -log \frac{minpath(c_1,c_2)}{2*D} \tag{3.2}$$

Nguyen et Al-Mubaid [117] ont incorporé à la fois la profondeur et le LCS dans leur mesure. Dans cette mesure, la similarité est le log de deux plus le produit de la distance la plus courte entre les deux concepts moins un et la profondeur de la taxonomie (D) moins la profondeur des concepts 'LCS (d) tel que défini dans Equation (3.3). Son étendue dépend de la profondeur de la taxonomie.

$$sim_{nam} = log(2 + (minpath(c_1, c_2) - 1) * (D - d) \tag{3.3}$$

B. Basée sur Le contenu informatif (IC-based)

Le contenu de l'information (IC) est formellement défini comme le log négatif de la probabilité d'un concept. Resnik [118] a modifié IC pour l'utiliser comme mesure de similarité. Il a défini la similarité de deux concepts comme étant le CI de leur LCS, comme le montre l'équation (3.4) :

$$sim_{res} = IC\big(lcs(c_1, c_2)\big) = -log(P\big(lcs(c_1, c_2)\big)) \tag{3.4}$$

Jiang et Conrath [119] et Lin [120] ont étendu la mesure basée sur IC de Resnik en incorporant IC des concepts individuels. Lin [120] a défini la similarité entre deux concepts en prenant le quotient entre deux fois le CI des concepts du LCS et la somme du CI des deux concepts, comme indiqué dans l'équation (3.5). Ceci est similaire à la mesure proposée par Wu et Palmer; différant dans l'utilisation de IC plutôt que dans la profondeur des concepts.

$$sim_{lin} = \frac{2*IC\big(lcs(c_1,c_2)\big)}{IC(c_1)+IC(c_2)} \tag{3.5}$$

Jiang et Conrath ont défini la distance entre deux concepts comme étant la somme de l'IC des deux concepts moins le double de l'IC des concepts. Nous modifions cette mesure pour renvoyer un score de similarité en prenant l'inverse de la distance comme indiqué dans l'équation (3.6).

$$sim_{lin} = \frac{1}{IC(c_1)+IC(c_2)-2*IC\big(lcs(c_1,c_2)\big)} \tag{3.6}$$

6.2. Contenu de l'information

A. Basée sur Corpus

Le contenu de l'information est défini comme le log négatif de la probabilité d'un concept. Nous définissons la probabilité d'un concept c en additionnant la probabilité que le concept P (c) apparaisse dans un texte plus la probabilité que ses descendants P (d) se produisent dans le même texte que celui présenté dans l'équation (3.7).

$$P(c*) = P(c) + \sum_{d \in descendant(c)} P(d) \tag{3.7}$$

La probabilité initiale d'un concept P (c) et de ses descendants P (d) est obtenue en divisant le nombre de fois qu'un concept est vu dans le corpus, freq (d), par le nombre total de concepts, N, comme on le voit dans l'équation (3.8).

$$P(d) = freq(d)/N \tag{3.8}$$

B. Basée sur la taxonomie:

Le défi des calculs de probabilité pour les concepts est qu'un grand nombre d'annotations sont nécessaires, afin de fournir une couverture suffisante de la taxonomie sous-jacente pour obtenir des estimations raisonnables. Le contenu intrinsèque de l'information cherche à résoudre ce problème tout en capturant la généralité et le caractère concret d'un concept. Il évalue le caractère informatif du concept en fonction de son placement dans la hiérarchie en examinant ses liens entrants (ancêtres) et sortants (descendant). Dans ce travail, nous utilisons le calcul IC intrinsèque proposé par Sanchez et *al.* [121] défini dans l'équation (3.9).

$$IC(c) = -log\left(\frac{\frac{|leaves(s)|}{|subsumers(c)|}+1}{max\_leaves+1}\right) \tag{3.9}$$

Où «leaves» sont le nombre de descendants de concept c qui sont des nœuds feuilles, les subsumant sont le nombre d'ancêtres de concept c et max_leaves est le nombre total de nœuds feuilles de la taxonomie.

6.3. Mesures de Connexité (Relatedness measures)

Lesk [122] a introduit une mesure qui détermine la connexité entre deux concepts en comptant le nombre de chevauchements entre leurs deux définitions. Un chevauchement (overlap) est la plus longue séquence d'un ou plusieurs mots consécutifs qui apparaissent dans les deux définitions. Lors de la mise en œuvre de cette mesure dans WordNet, Banerjee et Pedersen [123] ont constaté que les définitions étaient courtes et ne contenaient pas suffisamment de chevauchements

pour distinguer plusieurs concepts. Ils ont donc étendu cette mesure en incluant la définition des concepts associés.

Patwardhan et Pedersen [124] ont étendu cette mesure en utilisant des vecteurs de cooccurrence de second ordre. Dans cette méthode, un vecteur est créé pour chaque mot dans la définition de concepts contenant des mots qui co-apparaissent avec lui dans un corpus. La moyenne de ces vecteurs de mots permet de créer un seul vecteur de cooccurrence pour le concept. La similarité entre les concepts est calculée en prenant le cosinus entre les vecteurs de second ordre des concepts.

7. Désambiguïsation et l'impact de la distance des termes du contexte

Dans cette section, nous présentons une méthode utilisée pour déterminer le concept du mot ambigu le plus approprié au contexte et évaluons également l'effet de la taille de la fenêtre de contexte sur cette méthode, en utilisant le degré de similarité sémantique et de connexité entre les concepts possibles et les termes entourant le mot ambigu.

Cette section est organisée comme suit : nous décrivons les méthodes utilisées pour la résolution de l'ambiguïté des termes, en utilisant des mesures de similarité sémantique et de connexité pour WSD. Ensuite, nous illustrons un exemple sous une étude. Après, nous présentons les données et le système d'évaluation, puis nous discutons les résultats de l'évaluation de ces méthodes de WSD et finalisé par une synthèse.

7.1. SenseRelate

Cet algorithme est une implémentation k pour mesurer « gloss overlap » [122]. L'algorithme tente d'identifier la signification la plus probable d'un mot dans un contexte donné en se basant sur les relations sémantiques entre le mot cible et ses mots voisins dans une phrase.

Pour mettre en œuvre l'algorithme SenseRelate, nous procédons comme suit: nous commençons par l'identification des termes entourant le mot ambigu à l'aide du lexique SPECIALIST. Par la suite, nous calculons la corrélation entre le concept possible du mot ambigu et chacun des termes qui l'entourent en utilisant le package Perl UMLS: Similarité [126].

L'algorithme prend en compte le poids de la distance qui sépare les termes du contexte au mot cible, en multipliant l'inverse de cette distance par le score de similarité renvoyé. Cela peut avoir un impact négatif sur les concepts ou les sens cédés.

Pour surmonter ce problème et donc pour améliorer le processus de Biomedical WSD, nous proposons dans ce travail, une version modifiée simple de l'algorithme SenseRelate nommée No-DistanceSenseRelate, qui ignore simplement la distance, c'est-à-dire, les termes dans le contexte auront le même poids de distance.

7.2. NoDistanceSenseRelate

```
function getConceptNoDistanceSenseRelate (concepts, typeRelatedness, termsWindow)
return Concecpt
    score <- 0;
    conceptPref <- null;
    for each concept in concepts do
        sum <- 1;
        nbrTerms <- 0;
        for each term in termsWindow do
            maxScore <- 0;
            for each conceptUi in termsWindow->getConcepts() do
                similarity <- getRelatednessMeasure(concept,conceptUi, typeRelatedness);
                    if similarity > 0 then
                        maxScore <- maxScore + similarit;
                end if
            end for
            if maxScore > 0 then
                sum <- sum + maxScore  [For SenseRelarte -> maxScore * (1/term->getDist())]
                nbrTerms <- nbrTerms + 1;
            end if
        end for
        sum <- sum / nbrTerms;
        if sum > score then
            score <- sum;
            conceptPref <- concept;
        end if
    end for
    return conceptPref;
end function
```

Fig 5 Algorithme SenseRelate et NoDistanceSenseRelate

L'algorithme NoDistanceSenseRelate est une version modifiée de SenseRelate, dont nous ne multiplions pas le score de similarité par la distance entre le terme entourant le mot ambigu et le mot cible. Le score n'est donc pas basé sur la distance qui le sépare du mot cible. Dans la section suivante, nous donnons un exemple en illustrant les méthodes SensRelate et NoDistanceSenseRelate.

Le pseudocode des algorithmes SenseRelate et NoDistanceSenseRelate est décrit par l'algorithme dans la figure 5.

7.3. Etude de cas - Exemple

Pour donner un exemple, considérons la phrase suivante contenant le mot cible «Yellow fever» qui contient les concepts possibles : la fièvre jaune [C0043395] et le vaccin contre la fièvre jaune [C0301508].

Dans cet exemple, nous utilisons une taille de la fenêtre faisant référence aux trois termes de contenu situés à droite et à gauche du mot cible et tentons de les mapper sur des CUI. Dans ce cas, les mots contenus sont : « arthropod », « human », « west nile », « secretin test », « tick ».

Les algorithmes WSD (SenseRelate et NoDistanceSenseRelate) obtiennent ensuite des scores de similarité entre chacun des concepts possibles et les concepts des mots dans la fenêtre du contexte. Pour la méthode SenseRelate, chaque score est ensuite multiplié par l'inverse de sa distance par rapport aux mots cibles. Par exemple, arthropode est situé loin par trois mots de la fièvre jaune « Yellow Fever » et donc multiplié par 1/3. Mais pour NoDistanceSenseRelate, nous ne multiplions pas par la réciproque de la distance. Les scores sont ensuite additionnés pour obtenir un score total pour chaque concept possible, comme indiqué dans la figure 6 et le tableau 4.

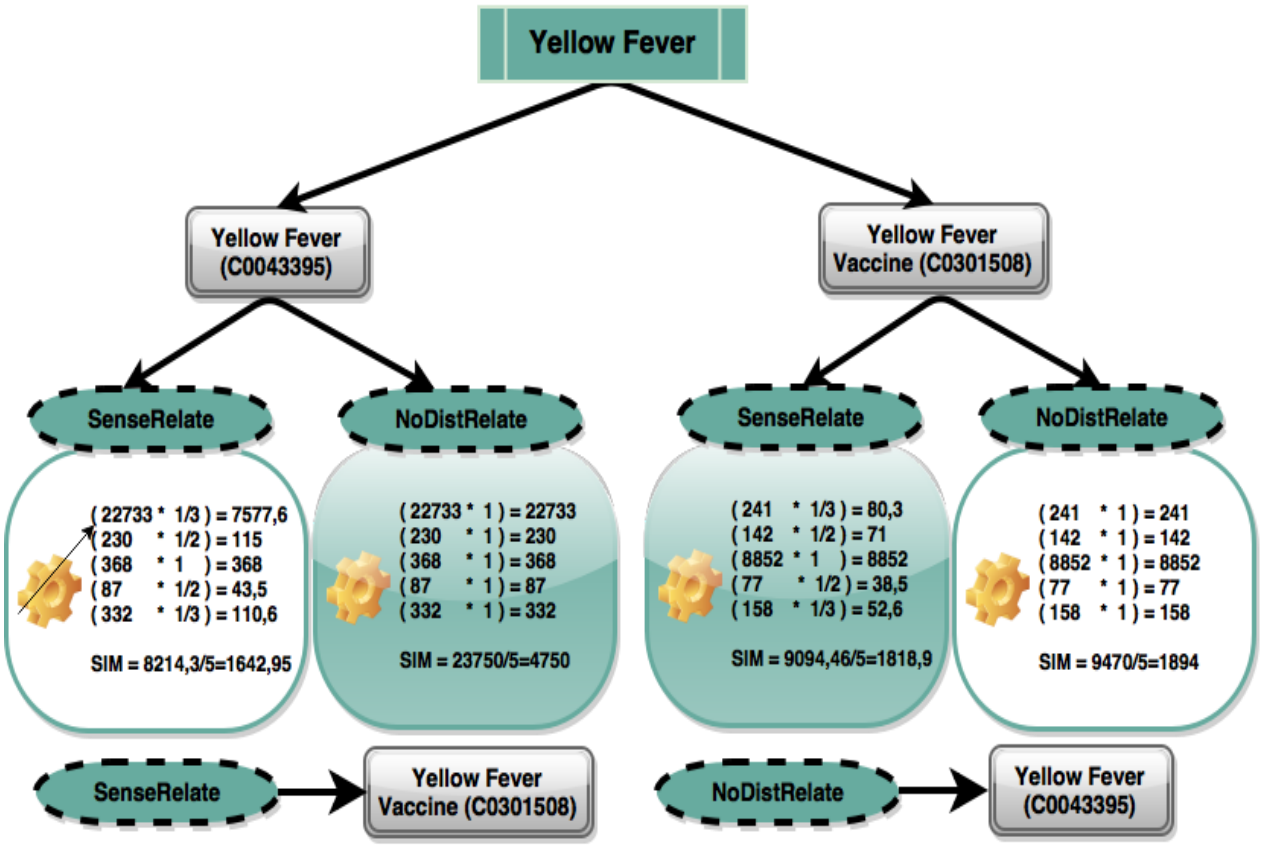


Fig 6 Exemple de méthode de calcul de WSD

Après avoir calculé le score de chaque méthode, nous déduisons le concept attribué à SenseeRelate et à NoDistanceSenseRelate. Dans cet exemple, les experts du MSH-WSD avaient choisi le concept C0043395, qui est celui sélectionné par NoDistanceSenseRelate.

Tableau 4 Exemple de mésure de connexité « Adapted Lesk »

Lesk measure	**Possible concepts**		**C0043395**	**C0301508**
	Label	***CUI***		
Arthropod	Arthropods	C0003903	339	145
	Poisoning due to venomous arthropod NOS	C0413111	105	70
	Arbovirus Infections	C0003723	22733	241
Human	Homo sapiens	C0086418	230	142
	Human Engineering	C0020120	100	79
	LCN2 protein, human	C0215955	71	65
West nile	West Nile Virus Vaccines	C1299874	132	8852
	West Nile Fever	C0043124	299	71
	West Nile virus	C0043125	368	164
secretin test	Secretin Measurement	C0523887	23	19
	Secretin stimulation test	C0430161	87	77
Tick	Ticks	C0040203	278	125
	Behavioral tic	C0278076	332	158
	Tick Paralysis	C0040197	28	8

7.4. Système d'évaluation

Dans cette section, nous présentons le Corpus et la plate-forme utilisée pour évaluer les méthodes SenseRelate et NoDistanceSenseRelate.

A. Corpus

Nous évaluons notre méthode sur le corpus MSH-WSD de NLM développé par Jimeno-Yepes et *al.* [125]. L'ensemble des données contient 203 termes ambigus et acronymes issus du référence 2010 de MEDLINE. Chaque instance d'un terme a été automatiquement attribuée un CUI de la version 2009AB de l'UMLS en exploitant le fait que chaque instance de MEDLINE est indexée manuellement avec des en-têtes de sujets médicaux dans lesquels chaque en-tête est associé à un CUI. Chaque mot cible contient environ 187 occurrences, 2,08 concepts possibles et 54,5% de sens majoritaire.

B. Plateforme

La figure 7 présente l'organigramme de notre plateforme d'évaluation. Comme définit dans la section précédente, nous utilisons MSH-WSD comme corpus pour évaluer les méthodes SenseRelate et NoDistSenseRealte. Nous commençons par l'extraction des termes pour le domaine biomédical afin d'extraire les termes biomédicaux du corpus en utilisant UMLS comme dictionnaire. Nous créons ensuite un objet Map qui contient le terme ambigu en tant que clé et la liste des termes dans le contexte en tant que valeur. Après nous mappons tous les termes sur leurs concepts en utilisant les services de terminologie UMLS (UTS) fournissent des services Web permettant de rechercher et de récupérer des données UMLS.

En résultant, la nouvelle Map contient des termes (termes ambigus et termes dans le contexte) avec leurs concepts, puis nous exécutons à la fois SenseRelate et NoDistSenseRealte (RelateAlgo) décrits dans la section 3, qui renvoient les concepts propres aux termes ambigus. Enfin, l'étape d'évaluation est exécutée afin de comparer les concepts sélectionnés à l'aide des méthodes et les concepts assignés par les experts du corpus MSH-WSD.

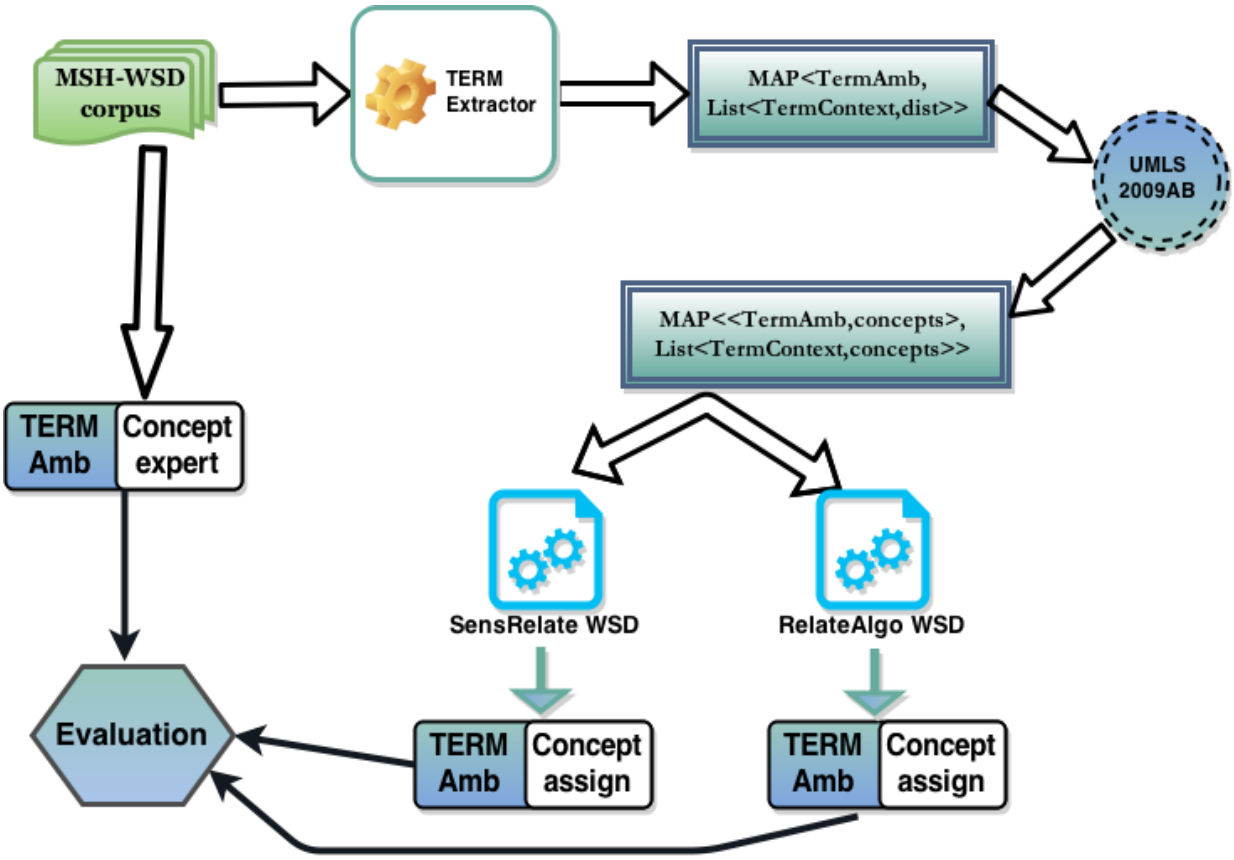


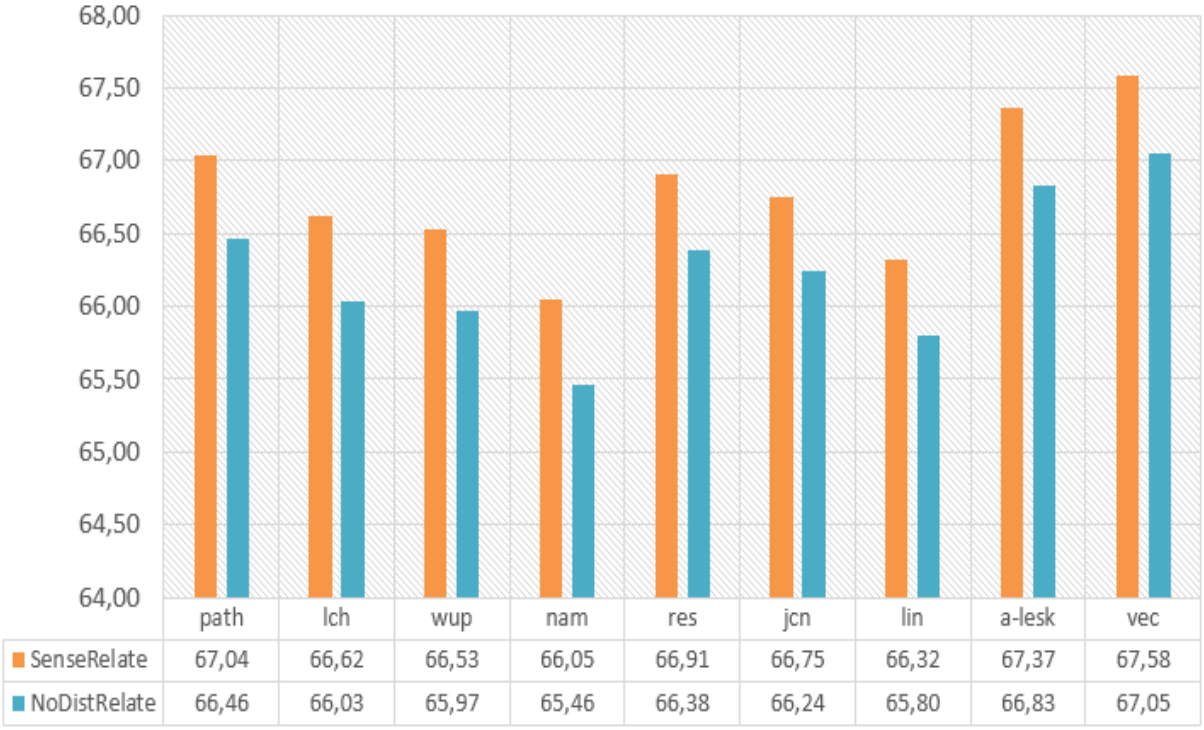
Fig 7 Organigramme de la Plateforme

7.5. Résultats et discussion

La figure 8 montre la précision des méthodes SenseRelate et NoDistRealte, en utilisant la mesure de chemin (Path) ; les mesures basées sur le chemin proposées par Leacock et Chodorow (Lch), Wu et Palmer (wup) et Nguyen et Al-Mubaid (nam); les mesures à base de CI proposées par Resnik (res), Jiang et Conrath (jcn) et Lin (lin) utilisant un contenu d'information basé sur la taxonomie; la mesure de parenté proposée par Banerjee et Pedersen (a-lesk); et la mesure de parenté proposée par Patwardhan et Pedersen (vec).

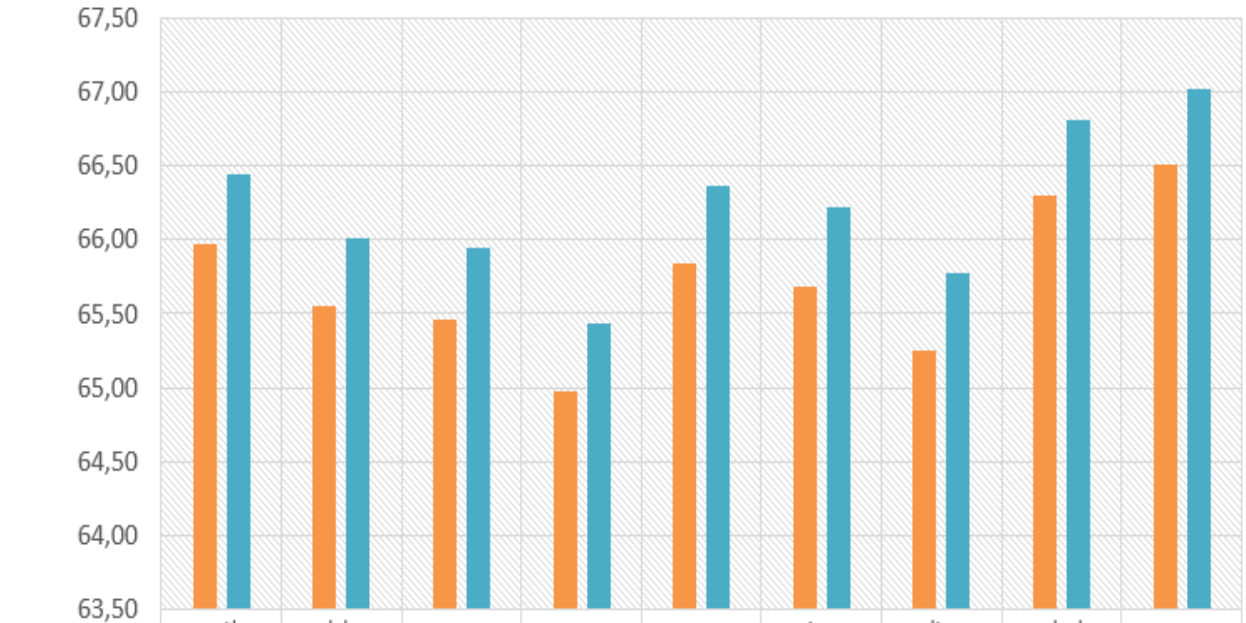
Les résultats sont affichés avec une taille de fenêtre de 2, ce qui signifie deux termes entourant le terme ambigu à droite et deux termes à gauche pour déterminer le concept approprié.

La figure 9 montre la précision des méthodes ***SenseRelate*** et ***NoDistSenseRealte***. Lorsque la taille de fenêtre est 3, cela signifie trois termes qui entourent le terme ambigu à droite et trois termes à gauche pour déterminer le concept approprié.



	path	lch	wup	nam	res	jcn	lin	a-lesk	vec
SenseRelate	67,04	66,62	66,53	66,05	66,91	66,75	66,32	67,37	67,58
NoDistRelate	66,46	66,03	65,97	65,46	66,38	66,24	65,80	66,83	67,05

Fig 8 Précision sur MSH-WSD en utilisant une similarité sémantique et une relation sur une fenêtre de taille 2



	path	lch	wup	nam	res	jcn	lin	a-lesk	vec
SenseRelate	65,97	65,55	65,46	64,98	65,84	65,68	65,25	66,30	66,51
NoDistRelate	66,44	66,01	65,95	65,44	66,36	66,22	65,78	66,81	67,02

Fig 9 Précision sur MSH-WSD en utilisant la similarité sémantique et la relation sur la taille de fenêtre de 3

Les résultats globaux pour les mesures de similarité sémantique et de connexité montrent que la méthode NoDistSenseRealte sur une fenêtre de taille 3 obtient toujours une précision de désambiguïsation supérieure à celle de SenseRelate, mais sur une fenêtre de taille 2, la méthode

SenseRelate obtient une précision de désambiguïsation supérieure à celle de NoDistSenseRealte. En plus, les résultats montrent que les mesures de connexité en utilisant les mesures a-lesk ou vec obtiennent une précision de désambiguïsation plus élevée que les autres mesures.

La figure 10 résume les résultats globaux en montrant la précision de SenseRelate et No-DistanceSenseRelate sur les fenêtres de tailles 2 et 3.

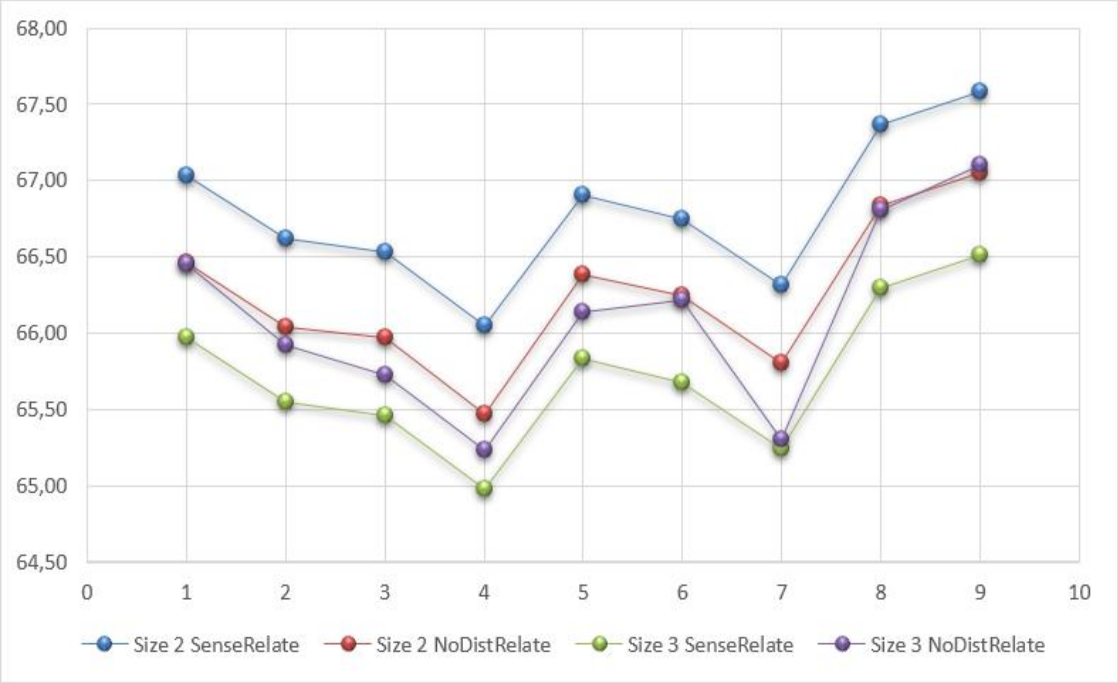


Fig 10 Précision sur MSH-WSD en utilisant la similarité sémantique et la parenté sur une taille de fenêtre de 2 et 3

8. Conclusion

Dans ce chapitre, nous avons étudié l'intégration de la sémantique dans la représentation par la conceptualisation. Nous avons choisi l'utilisation d'une première étape de reconnaissances des entités nommées biomédicales, avant d'entamer le processus de conceptualisation afin de tirer parti des résultats issues de notre étude concernant la Bio-NER présenté dans le chapitre précédent. Ainsi dans l'approche proposée, nous appliquons la conceptualisation aux termes issus de system de BioNER à travers l'enrichissement par les concepts de l'ontologie utilisée. Nous avons également discuté les différentes stratégies de conceptualisation et de résolution de l'ambiguïté et présenté un cadre de travail générique pour la conceptualisation de texte.

Ensuite, nous avons évalué une méthode appelée SenseRelate pour WSD. L'objectif de ce travail était d'évaluer l'influence de la taille de la fenêtre contextuelle sur cette méthode en utilisant le degré de similarité sémantique et la connexité extrapolée à partir de l'UMS. Pour ce faire, nous

avons apporté quelques modifications à SenseRelate de manière à ce que le score ne soit pas basé sur sa distance par rapport au mot cible, ce qui génère une nouvelle méthode appelée NoDistance-SenseRelate, et nous avons évalué SenseRelate et NoDistanceSenseRelate sur le corpus MSH-WSD en utilisant les mesures de similarité sémantique et de connexité du package UMLS::Similarity. D'après les résultats obtenus, nous avons constaté sur ce corpus que la méthode NoDistSenseRealte sur une fenêtre de taille 3 obtenait systématiquement une précision de la désambiguïsation supérieure à celle de SenseRelate. Cependant, pour une taille de fenêtre de 2, la méthode SenseRelate obtenait une précision de la désambiguïsation supérieure à celle de NoDist-SenseRealte. Nous concluons donc que la distance entre le mot cible et les termes du contexte peuvent influencer négativement ou positivement le score final, en fonction de la taille de la fenêtre.

Chapitre 4 : Enrichissement Sémantique du Texte Court : Web-Kernel

1. Introduction

Les algorithmes de WSD identifient automatiquement le sens approprié d'un mot ambigu en fonction du contexte dans lequel le mot est utilisé. Pour la désambiguïsation des sens basée sur la connaissance non supervisée, afin de déterminer le sens correct (concept correct), on utilise généralement des mesures de similarité sémantique ou de connexité [65] qui tentent de quantifier la proximité sémantique entre deux concepts du contexte. Les mesures de Connexité (Relatedness measures) déterminent la connexité entre deux concepts en se basant sur leurs définitions, en raison de la définition abrégée d'un concept, nous pensons qu'une similarité de texte abrégée peut être utile dans le processus du WSD pour le domaine Biomédicale. Dans ce chapitre, nous proposons une nouvelle mesure de similarité au niveau du noyau Web (Web-kernel) entre une paire de textes courts biomédicaux, dérivés de la définition de concepts et le résultat de recherche du moteur de recherche PubMed basée sur la théorie des ensembles approximatifs (RST).

Pour calculer la corrélation entre une paire de textes courts biomédicaux, la mesure proposée étend la fonction de similarité Web-Kernel introduite par Sahami et Heilman [127], où ils exploitent les résultats de la recherche sur le Web pour fournir un contexte plus large aux textes courts. Dans notre contribution, nous enrichissons la représentation des résultats de recherche sur le Web en se basant sur la théorie des ensembles approximatifs, qui est un outil mathématique permettant de traiter le manque de précision et l'incertitude [128]. D'autre part, nous considérons un nouveau schéma de pondération au lieu d'utiliser TF * IDF [129].

Ce chapitre est organisé comme suit, nous présentons l'arrière-plan mathématique de la théorie des ensembles bruts et notre mesure de similarité améliorée pour les textes courts, ensuite nous décrivons la méthode utilisées pour la résolution de l'ambiguïté des termes basés sur la mesure de similarité Web-kernel. Après nous présentons les données et le système d'évaluation, puis nous discutons des résultats de l'évaluation de cette méthode et une synthèse. La dernière section est consacrée pour des remarques finales

2. Théorie des Ensembles Approximatifs

2.1. Approximation d'un Concept

La théorie des ensembles approximatifs a été développée à l'origine par un informaticien polonais Zdzisław Pawlak [128] en tant qu'un outil d'analyse et de classification des données. Elle a été appliquée avec succès dans diverses tâches, telles que la sélection / extraction de caractéris-

tiques et la classification [128] [130]. Dans ce chapitre, nous présentons les concepts fondamentaux des ensembles approximatifs avec des exemples illustratifs. On outre, nous décrivons le modèle de tolérance dédié aux traitements de données textuelles.

Considérons un ensemble d'objet non vide U appelé l'univers. Supposons qu'on veut définir un concept sur l'univers des objets U ; et que notre concept peut être représenté comme un sous-ensemble X de U. Le point central de la théorie des ensembles approximatifs est la notion d'approximation d'ensemble : Tous ensemble dans U peut être approché par son approximation inférieure et supérieure.

A. Relation d'indiscernabilité

A l'origine [130], pour définir une approximation inférieure et supérieure, il faut introduire une relation d'indiscernabilité : R □ U × U (où R peut être n'importe quelle relation d'équivalence : réflexive, symétrique et transitive). Pour deux objets x, y ∈ U, si xRy (c'est-à-dire qu'il y a une relation entre x et y) alors on dit que x et y sont indiscernables l'un de l'autre. La relation d'indiscernabilité R conduit à une partition complète de l'univers U en classes équivalentes [x]R, avec x ∈ U.

Nous définissons l'approximation inférieure et supérieure de l'ensemble X, par rapport à un espace d'approximation désigné par A = (U ; R), respectivement comme suit :

$$LR(X) = \{x \in U: [X]R \subseteq X\} \quad (4.1)$$

$$UR(X) = \{x \in U: [X]R \cap X \neq \emptyset\} \quad (4.2)$$

Intuitivement, nous pouvons déduire que l'approximation inférieure contient des objets qui appartiennent certainement à notre concept, tandis que l'approximation supérieure contient des objets qui peuvent appartenir à notre concept (Figure 11).

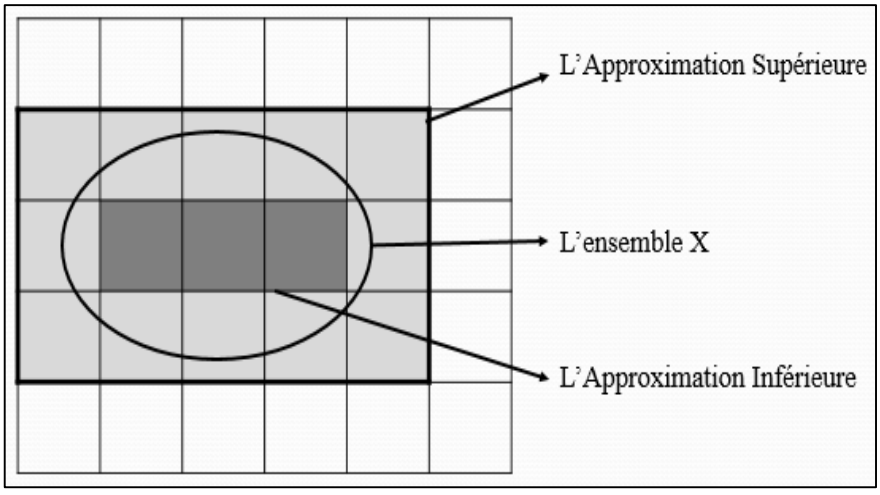


Fig 11 Visualisation de l'approximation inférieure et supérieure d'un ensemble X

B. Fonction d'Appartenance

Les approximations peuvent aussi être définies par la fonction d'appartenance approximative [95]. Étant donné la fonction d'appartenance approximative μX : U → [0, 1] d'un ensemble X □ U, les approximations inférieure et supérieure sont définies comme suit :

$$L\mu(X) = \{x \in U: \mu(x, X) = 1\} \quad (4.3)$$

$$U\mu(X) = \{x \in U: \mu(x, X) > 0\} \quad (4.4)$$

Noter que, la fonction d'appartenance approximative est définie comme μ(x, X) = |[X] R ∩ X| / |[X] R|, cette définition est en réalité équivalente à (4.1) et (4.2) respectivement.

C. Le Système d'Information

Un système d'information est décrit comme une paire (U, A) ; où U est un ensemble fini et non vide d'objets appelé l'univers, A est un ensemble fini non vide des attributs qui décrivent les objets. Pour tout a ∈ A correspond une fonction fa : U → Va (où Va est l'ensemble des valeurs de a). Généralement, un système d'information est représenté par une table à deux dimensions où les lignes représentent les objets et les colonnes représentent les attributs.

D. Exemple Illustratif

Le tableau 5 présente un exemple de système d'information d'un univers de patients hospitalisés U = {p1, p2, p3, p4, p5, p6, p7, p8}. Les patients sont décrits par leurs symptômes de la maladie. L'ensemble des attributs est définie comme : A= {Mal à la tête, Douleur Musculaire, Température}, avec :

- $V_{\text{Mal à la tête}}$ = {Oui, Non}
- $V_{\text{Douleur Musculaire}}$ = {Oui, Non}
- $V_{\text{Température}}$ = {Normal, Élevé, Trop Élevé}

Tableau 5 Exemple d'un Système d'Information

Patient	**Attributs**		
	Mal à la tête	**Douleur Musculaire**	**Température**
p1	Oui	Oui	Normal
p2	Oui	Oui	Élevé
p3	Oui	Oui	Trop Élevé
p4	Non	Oui	Normal

p5	Non	Non	Élevé
p6	Non	Oui	Trop Élevé
p7	Non	Non	Élevé
p8	Non	Oui	Trop Élevé

Généralement, le tableau d'information est donné avec une colonne additionnelle d dite colonne de décision, dans ce cas, nous parlons d'un Système de Décision : I = (U, A U {d}). Au système d'information, on va ajouter une nouvelle colonne de décision {grippe} avec : V grippe = {Oui, Non}. Voici la nouvelle représentation de notre système de décision (Tableau 6).

Soient une relation d'équivalence R définie par l'égalité des deux attributs 'Mal à la tête' et 'Température' (xRy ⟺ f Mal à la tête (x) = f Mal à la tête (y) ^ f Température(x) = f Température(y)). La relation d'équivalence R partitionne l'univers U en des classes dont les éléments partagent les mêmes propriétés. Voici les classes obtenues : {p1}, {p2}, {p3}, {p4}, {p5, p7} et {p6, p8}. Nous pouvons voir par exemple, que p5 et p7 sont indiscernables à l'égard de relation R.

Tableau 6 Exemple d'un Système de Décision

Patient	**Attributs**			**Grippe**
	Mal à la tête	**Douleur Musculaire**	**Température**	
p1	Oui	Oui	Normal	Non
p2	Oui	Oui	Élevé	Oui
p3	Oui	Oui	Trop Élevé	Oui
p4	Non	Oui	Normal	Non
p5	Non	Non	Élevé	Non
p6	Non	Oui	Trop Élevé	Oui
p7	Non	Non	Élevé	Oui
p8	Non	Oui	Trop Élevé	Non

Supposant qu'un concept X que nous voulons décrire est donné ici comme une décision, si un patient a la grippe. Dans ce cas, le concept X sera défini comme suit : X = {p2, p3, p6, p7}. Pour déduire les approximations de X, il faut calculer pour chaque objet de U sa classe d'équivalence suivant la relation R fixée précédemment. Voici les classes d'équivalence obtenues :

- $[p1]_R = \{p1\}$
- $[p2]_R = \{p2\}$
- $[p3]_R = \{p3\}$
- $[p4]_R = \{p4\}$
- $[p5]_R = \{p5, p7\}$
- $[p6]_R = \{p6, p8\}$
- $[p7]_R = \{p5, p7\}$
- $[p8]_R = \{p6, p8\}$

Finalement, les approximations inférieure et supérieure de X selon l'équation (4.1) seront :

$L_R(X) = \{p2, p3\}$
$U_R(X) = \{p2, p3, p5, p6, p7, p8\}$

Pour certaines applications, l'exigence d'une relation équivalente s'est révélée trop stricte. Par conséquent, un espace généralisé de tolérance est introduit par Skowron [131], permettant de passer d'une relation d'équivalence à une autre relation, dite de tolérance, où la propriété de transitivité n'est plus requise.

2.2. Modèle de Tolérance des Ensembles Approximatifs (Rough Set Tolerance Model - RSTM)

Le modèle de tolérance des ensembles approximatifs a été mis au point [132] [133] pour modéliser les documents et les termes dans la recherche d'information, fouille de textes, etc. Avec sa capacité de gérer l'incertitude et l'imprécision, le modèle de tolérance paraît être un bon outil pour modéliser la relation entre les termes et les documents.

A. L'Espace de Tolérance des Termes (Tolerance Space)

Nous considérons D = {d1, d2, …. dn} un ensemble de documents, T = {t1, t2, …. tm} est l'ensemble des termes de D. En adoptant le modèle d'espace vectoriel, chaque document sera représenté par un vecteur de poids [wi1, wi2, …. wij], où wij représente le poids du terme j dans le document i. Dans le modèle de tolérance, l'espace de tolérance est défini sur un univers de tous les termes de la façon suivante : U = T = {t1, t2, …. tm}.

L'idée est de reproduire les termes indices qui sont reliés sémantiquement dans des classes. Pour ce faire, la relation de tolérance est définie comme la cooccurrence des termes indices dans

tous les documents de D. Le choix de cette relation est motivé par sa capacité de prendre en considération la relation sémantique entre les termes. C'est-à-dire, plus deux termes se reproduisent ensemble dans plusieurs documents, plus ces deux termes sont liés sémantiquement.

B. Classe de Tolérance d'un Terme (Tolerance Class)

Nous considérons $f_D(ti, tj)$ le nombre de documents dans D, dans lesquels apparaitront les deux termes ti et tj. La fonction d'incertitude I avec un seuil θ est définie par :

$$I\theta(ti) = \{tj \mid fD(ti, tj) > \theta\} \cup \{ti\} \quad (4.5)$$

Clairement, nous pouvons déduire que la fonction (4.5) satisfait aux deux conditions d'être réflexive ($ti \in I\theta(ti)$) et symétrique ($ti \in I\theta(tj) \Leftrightarrow tj \in I\theta(ti)$ pour tout $ti, tj \in T$). Donc, Iθ (ti) est la classe de tolérance du terme index ti.

Une classe de tolérance représente un concept caractérisé par les termes qu'il contient. En faisant varier le seuil θ (par exemple par rapport à la taille de la collection de documents), nous pouvons contrôler le degré de liaisons des mots dans les classes de tolérance.

La fonction d'appartenance μ pout tout $ti \in T$, $X \subseteq T$ est définie par :

$$\mu X(ti, X) = v(I\theta(ti), X) = |I\theta(ti) \cap X| / |I\theta(ti)| \quad (4.6)$$

Finalement, les deux approximations inférieure et supérieure de tout sous-ensemble X □ T peuvent être déterminées de la façon suivante :

$$LR(X) = \{ti \in T : v(I\theta(ti), X) = 1\} \quad (4.7)$$

$$UR(X) = \{ti \in T : v(I\theta(ti), X) > 0\} \quad (4.8)$$

Si nous considèrons X comme un concept qui est décrit d'une façon vague par les termes qui les constitue, dans ce cas nous pouvons dire que l'approximation supérieure est l'ensemble des termes qui partagent des liens sémantique avec X, alors que l'approximation inférieure représente le corps du concept X.

C. Enrichissement de la Représentation d'un Document

Dans le modèle d'espace vectoriel (Vector Space Model), un document est vu comme un sac de termes (mots) pondérés ; ceci est représenté en assignant des valeurs de poids, dans le vecteur de document, différentes de zéro pour les termes qui se produisent dans le document [134]. Avec le modèle de tolérance des ensembles approximatifs, l'objectif est d'enrichir la représentation du document en tenant compte non seulement aux termes du document, mais également d'autres termes avec lesquels existent des liens sémantiques.

Une représentation plus «riche» du document peut être acquise en le représentant comme un ensemble des classes de tolérance des termes qu'il contient. Ceci, est réalisé simplement en représentant le document avec son approximation supérieure. Soit : di = {ti1, ti2, …. tik} un document dans l'ensemble des documents D, avec ti1, ti2, …. tik ses termes ; alors l'approximation supérieure du document di est :

$$UR\ (di) = \{ti \in T : v\ (I\theta\ (ti), di) > 0\} \tag{4.9}$$

D. Schéma de Pondération Etendu pour l'Approximation Supérieure

Pour construire le vecteur de poids du document, il suffit d'utiliser la méthode TF*IDF, mais afin d'utiliser les approximations du document, le système de pondération doit faire la différence entre les termes qui existent juste dans son approximation supérieure et non pas dans le document lui-même (c'est-à-dire les termes ajoutés après le calcul de l'approximation supérieure). Le système de pondération est défini à travers l'équation (4.10) comme [129] :

$$wij = \begin{cases} (1 + \ln(f_{di}(t_i))) * \ln\frac{N}{f_D(t_j)} & ; t_j \in d_i \\ min\ tk \in wij * \frac{ln(\frac{N}{f_D(t_j)})}{1+ln(\frac{N}{f_D(t_j)})} & ; tj \in UR(di) / d \\ 0 & ; t_j \notin U_R(d_i) \end{cases} \tag{4.10}$$

Où wij est le poids de terme tj dans le document di. L'extension veille à ce que les termes apparaissant dans l'approximation supérieure de di mais pas en di, ont un poids plus petit que le poids des termes du document di. Après le calcul des poids avec le système de pondération étendu, une normalisation par la longueur de vecteur est ensuite appliquée à tous les vecteurs des documents en utilisant l'équation (4.11) [129] :

$$wij = \frac{w_{ij}}{\sqrt{\sum_{t_k \epsilon U_R(d_i)} (w_{ij})^2}} \tag{4.11}$$

3. Mesure de Similarité pour le Texte Court

Les algorithmes de WSD identifient automatiquement le sens approprié d'un mot ambigu en fonction du contexte dans lequel le mot est utilisé, ces algorithmes utilisent généralement des mesures de similarité sémantique ou de connexité [65] qui tentent de quantifier la proximité sémantique entre deux concepts du contexte. Les mesures de connexité (Relatedness measures) déterminent la connexité entre deux concepts en se basant sur leurs définitions, en raison de la définition

abrégée d'un concept, nous pensons qu'une similarité de texte abrégé peut être utile dans le processus du WSD pour le domaine biomédical.

Mesurer la similarité de textes courts est devenue une étape cruciale dans le processus de nombreuses tâches de fouille de texte en raison de leurs brièvetés et le manque d'information contextuelle. Par conséquent, elle a été étudiée d'une manière intensive durant la dernière décennie.

Trois approches ont été utilisées dans la littérature pour mesurer la similarité entre deux textes :

3.1. Approche basée sur les Mots Partagés (Surface Matching)

Étant donné une paire de texte (x, y), l'idée est basée sur le nombre de mots qui apparaissent dans les deux textes. Cette technique repose sur l'hypothèse que les textes similaires partagent de mots identiques, ce qui n'est efficace que si les deux textes ont une longueur suffisante. Disons que X et Y sont les ensembles de mots dans les deux textes x et y. Les mesures de similarité présentées dans [135] sont listées comme suit :

- Coefficient de correspondance simple (Simple Matching) qui est le nombre de termes partagés. Ce coefficient ne prend pas en compte les tailles de X et Y. Il a été utilisé dans la recherche d'information et peut être défini comme suit :

$$Mat(x, y) = |X \cap Y| \tag{4.12}$$

- Coefficient de similarité Jaccard (Jaccad Measure) est une mesure statistique utilisée pour comparer la similarité et la diversité des ensembles d'échantillons. Le coefficient entre les ensembles d'échantillons est défini comme la taille de l'intersection divisée par la taille de l'union des ensembles d'échantillons :

$$Jacc(x, y) = |X \cap Y| / |X \cup Y| \tag{4.13}$$

- Le coefficient de Dé (Dice Measure) est une mesure de similarité liée au coefficient Jaccard. Il peut être défini comme suit :

$$Dice(x, y) = |X \cap Y| / (|X| \cap |Y|) \tag{4.14}$$

- Le coefficient de chevauchement (Overlap Measure) est une mesure de similarité qui calcule le chevauchement entre deux ensembles. Si l'ensemble X est un sous-ensemble de Y ou l'inverse. Il est défini comme suit :

$$Over(x, y) = |X \cap Y| / min\ (|X|, |Y|) \tag{4.15}$$

- Le Coefficient Cosinus est une mesure de similarité entre deux vecteurs de n dimensions en trouvant l'angle entre eux. Il est souvent utilisé pour comparer des documents dans la recherche d'information et le regroupement de texte. Il peut être défini comme :

$$Cos(X, Y) = |X \cap Y| / \sqrt{|X| * |Y|} \quad (4.16)$$

Ces techniques ne sont efficaces et adéquates que si on compare des textes comportant un nombre de mots suffisant. Avec des textes courts, ces techniques présenteraient des inconvénients évidents. Par exemple, deux phrases peuvent avoir un sens similaire mais elles sont composées de mots totalement différents ; ou même dans les cas où deux textes courts peuvent partager des termes, ils peuvent être utilisés dans des contextes différents. Il existe également plusieurs extensions possibles de ces méthodes. Par exemple, deux termes peuvent être équivalents s'ils sont synonymes selon certains thésaurus tels que WordNet; ou bien les ensembles X et Y peuvent être construits en utilisant les racines ou les stems au lieu des termes originaux.

3.2. Approche basée sur l'Historique des Requêtes (Query log Methods)

++* la suite utilisées pour la génération de suggestions de requêtes de recherche, qui devient de plus en plus précise et utile [136]. Un exemple de substitution de requête est présenté dans [137] où Jones et *al.* ont introduit la notion de substitution de requête, c'est-à-dire la génération d'une nouvelle requête pour remplacer la requête de recherche initiale d'un utilisateur. Ils ont utilisé des modifications basées sur les substitutions habituelles apportées par les internautes à leurs requêtes. De cette manière, la nouvelle requête est fortement liée à la requête d'origine et devra contenir des termes fortement liés à tous les termes d'origine.

3.3. Approche basée sur les Corpus (Corpus based Methods)

Il existe deux types de méthodes populaires basées sur les corpus. Le premier type suppose que la signification d'un mot peut être définie à partir de son utilisation (c'est-à-dire sa distribution dans le texte); L'analyse sémantique latente (Latent Semantic Analysis - LSA) [138], l'analyse sémantique probabiliste latente (Probabilistic Latent Semantic Analysis - PLSA) [139]. Par la suite l'allocation de Dirichlet latent (Latent Dirichlet Allocation - LDA) [140] sont les méthodes les plus connues. Le second type basé sur les corpus utilise des connaissances externes telles que WordNet ou le corpus Web pour calculer la similarité. Cela ne convient pas aux langues qui ne disposent pas de ressources ou des bases de connaissance telles que WordNet ou UMLS.

Récemment, Adel et *al.* [141] ont proposé FUSE (FUzzy Similarity mEasure) une mesure de similarité pour le texte court basée sur un algorithme de la logique flou. Les auteurs ont testé et évalué en utilisant trois corpus, les résultats obtenus montrent l'intérêt de leur proposition. Nagoudi et Schwab [142] ont proposé un système basé sur les termes incorporés (Words Embedding) pour calculer la similarité entre deux textes courts. L'idée principale est d'exploiter les vecteurs en tant que représentations de mots dans un espace multidimensionnel afin de dégager les propriétés

sémantiques et syntaxiques des mots. Un analyseur morphosyntaxique et un système de pondération TF-IDF sont appliqués pour améliorer l'identification des mots les plus descriptifs d'un texte court. Nakamura et *al.* [143] ont utilisé Wikipédia pour calculer la similarité entre les textes courts multilingues. Shirakawa et *al.* [144] ont proposé d'utiliser une analyse probabiliste pour calculer les poids des termes clés et par la suite, trouver les entités associées à chaque terme clé dans Wikipédia. Les termes clés et les vecteurs des entités extraites seront utilisés dans une version étendue du théorème de Naïve Bayes.

Les méthodes de l'approche basée sur les corpus ont prouvé leur efficacité par rapport aux deux premières approches qui sont basées respectivement sur Les Mots Partagés et l'Historique des requêtes. Parmi les méthodes les plus connues de la littérature, nous trouvons celle introduite par Sahami et Heilman [145] pour mesurer la similarité entre des textes courts en exploitant les résultats retournés par un moteur de recherche Web pour fournir un contexte plus large au contenu initial. Considérant une paire de texte court, ils ont traité chaque élément comme une requête à envoyer vers un moteur de recherche Web, puis à partir des résultats renvoyés, ils ont créé pour chaque requête un vecteur de contexte contenant de nombreux mots ayant tendance à apparaître dans le contexte des termes d'origine, en utilisant TF * IDF comme système de pondération. Enfin, la similarité est définie comme le produit scalaire des deux vecteurs de contexte pour les deux textes courts.

Soit un texte court X, le vecteur de contexte de x noté QE(x) (Query Expansion) sera calculé de la façon suivante :

1. Envoyer x en tant que requête auprès un moteur de recherche.
2. Soit R (x) l'ensemble des n top documents récupérés d_1; d_2; …; d_n
3. Calculer le vecteur de terme TF-IDF v_i pour chaque document $d_i \in R(x)$
4. Soit C(x) le centre des vecteurs v_i selon la norme L_2 : $C(x) = \sum_{i=1, n} v_i / \| v_i \|$
5. Soit QE(x) la normalisation centre de gravité C (x) : $QE(x) = C(x) / \| C(x) \|$

Considérant deux textes courts x et y, la similarité entre les deux peut être définie comme étant le produit scalaire des deux vecteurs de contexte : QE(x).QE(y)

Cette méthode peut être améliorée en apportant quelques traitement aux documents retournés par le moteur de recherche dans l'étape 2, ou bien en adoptant un nouveau système de pondération dans l'étape 3 au lieu du système utilisé TI-IDF. Nous présenterons une version améliorée de cette mesure dans la section suivante.

4. Méthode proposée

Dans cette section, nous présentons en détail notre méthode proposée pour calculer la mesure de similarité appelée mesure de similarité Web-RST (Web-RST similarity measure) pour les textes courts biomédicaux (Figure 12). Dans notre proposition, nous avons deux contributions majeures par rapport à l'algorithme original présenté précédemment [127].

À l'étape 1, et en raison de la brièveté (Shortness) et de le manque de l'information contextuelle (Sparseness) des extraits renvoyés et qui sont générés pour représenter l'ensemble du document par un moteur de recherche, nous considérons l'approximation supérieure de chaque document (pouvant être calculée à l'aide de la formule (4.3)). Ce qui enrichie la représentation du document en ajoutant des termes qui ont certains liens sémantiques avec les termes originaux.

À l'étape 2, chaque document est représenté par un vecteur de pondération, la formule originale TF-IDF combinant les définitions de fréquence de terme et de fréquence inverse de document étant remplacée par la formule (4.17)

$$wij = \begin{cases} (1 + ln(f_{di}(t_i))) * ln\frac{N}{f_D(t_j)} \; ; \; t_j \epsilon \, d_i \\ min \; tk \epsilon \, wij * \frac{ln(\frac{N}{f_D(t_j)})}{1+ln(\frac{N}{f_D(t_j)})} \; ; \; tj \epsilon \, UR(di) \, / \, di \\ 0 \qquad\qquad\qquad ; \; t_j \notin U_R(d_i) \end{cases} \tag{4.17}$$

Où wij est le poids du terme j dans le document di. Cette formule garantit que chaque terme apparaissant dans l'approximation supérieure de di et non pas dans di a un poids inférieur à celui de tous les termes dans di. La normalisation par la longueur du vecteur est appliquée à tous les poids des vecteurs de document wij [129].

$$wij = \frac{w_{ij}}{\sqrt{\sum_{t_k \epsilon U_R(d_i)} (w_{ij})^2}} \tag{4.18}$$

Les étapes détaillées de la nouvelle méthode d'extension de la requête sont décrites ci-dessous :

1. Soit D_n (x) l'ensemble des n principaux documents renvoyés par un moteur de recherche lorsqu'on utilise x comme terme de requête (nous utilisons les services offerts par PubMed e-utilities [7]).

2. Chaque document en di ϵ D_n (x) sera nettoyé en supprimant les mots vides, les mots latins et les caractères spéciaux comme (/, #, $, etc…) et symbolisée. Ensuite, nous générons l'approximation supérieure U (di).
3. Pour chaque document U (di), nous construisons le terme vecteur vi, où chaque élément est le score pondéré w_{ij} du terme t_j, défini comme été montré dans la formule (4.17).
4. Soit C(x) le centre de gravité des vecteurs normalisés L2 vi: C (x) = vi / || vi ||
5. Soit QERST (x) la normalisation L2 du centroïde C (x): QERST (x) = C (x) / || C (x) ||
6. Avec une paire de segments de texte courts x et y, la mesure de similarité Web-Kernel est QERST (x) .QERST (y)

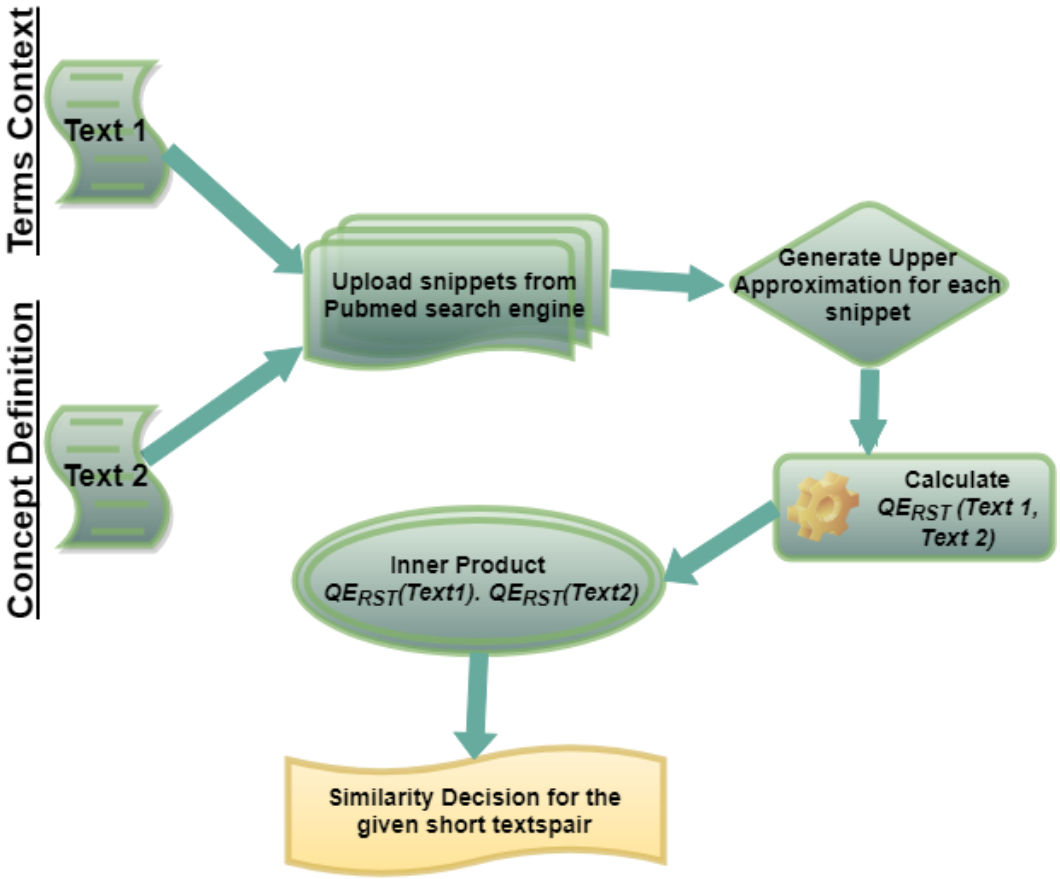


Fig 12 Organigramme de calcul de similarité

L'algorithme du processus de similarité Web-RST peut être résumé et présenté dans la figure 13.

Nous avons examiné plusieurs paires de texte courtes afin de déterminer le score de similarité attribué par le noyau basé sur le Web et notre méthode améliorée. Nous examinons en particulier trois genres de correspondance de texte court: les acronymes, les individus et leurs positions

et les termes à facettes multiples, ce qui signifie paire de textes courts contenant des termes pouvant avoir la même définition dans deux contextes, mais la sémantique acceptée de cette définition peut varier dans le contexte [146].

```
Algorithm: double WebRSTSim(String query1, String query2)
Begin
    list<Snippet> sl1 = NTopResults(query1)
    list<Snippet> sl2 = NTopResults(query2)
     preProcess(sl1)
     preProcess(sl2)
    generateUpperApproximation(sl1)
    generateUpperApproximation(sl2)
    Compute the term weight vector vi for each document di in sl1
    Compute the term weight vector wi for each document di in sl2

    //Let C1 be the centroid of the L2 normalized vectors vi
    for each vi do
        C1 = vi / ||vi||2
    end for
    C1 = C1 / N
    //Let QE1 the query expansion of query1
    QE1 = C1 / ||C1||2

    for each wi do
        C2 = wi / ||wi||2
    end for
    C2 = C2 / N
    QE2 = C2 / ||C2||2
    double sim = QE1 . QE2  // . is the inner product between QE1 and QE2
    return sim
End

void generateUpperApproximation(list<document> dl)
Begin
    Initialise the TermSet s
    Initialise the co-occurrence matrix
    for each term ti in s do
        for each term tj in s do
            if occurTogether(ti, tj) > θ then
                addToToleranceClass(ti, tj)
            end if
        end for
    end for
    for each document d in dl do
        for each term tj in s do
            tct = toleranceClass(tj)
            cof = |tct ∩ d| / |tct|
            if cof > 0 then
                addTermToupperApprox(d, tj)
            end if
        end for
    end for
end
```

Fig 13 Algorithme WebRSTSim

Dans cette section, nous avons présenté une méthode efficace pour calculer la similarité entre les paires de textes courts biomédicaux. Les résultats obtenus montrent l'intérêt de notre amélioration par rapport à la méthode d'origine. Cette méthode peut être utilisée comme une mesure de similarité de sens entre les concepts, comme présenté dans la section suivante.

5. Système d'évaluation

Afin d'évaluer la mesure de similarité du Web-kernel, nous développons un système de désambiguïsation du sens du mot basé sur cette méthode. Dans cette section, nous fournissons l'ensemble de données et l'organigramme du système d'évaluation.

5.1. Corpus

Nous évaluons la mesure de similarité entre Web-kernel dans le corpus MSH-WSD de NLM développé par Jimeno-Yepes et *al.* [147]. Le corpus comprend 106 abréviations ambiguës, 88 termes ambigus et 9, une combinaison des deux, pour un total de 203 mots ambigus. Chaque instance contenant le mot ambigu a été affectée à une CUI de la version 2009AB de l'UMLS. Pour chaque terme / abréviation ambigu, le jeu de données contient un maximum de 100 instances par sens obtenu à partir de MEDLINE; totalisant 37 888 cas d'ambiguïté sur 37 090 citations MEDLINE.

5.2. Organigramme du Système

Comme mentionné précédemment, nous utilisons MSH-WSD comme corpus pour évaluer la similarité du Web-kernel. La figure 14 présente l'organigramme de notre système d'évaluation.

Dans ce système, nous procédons comme suit :

- À partir du corpus, nous commençons par extraire le terme ambigu, puis exécuter le module *Term Extractor* pour le domaine biomédical, qui tente d'extraire des termes biomédicaux dans le contexte du mot ambigu. Considérant le terme ambigu: Paludisme de MSH-WSD: «Antimicrobiens chez les enfants hospitalisés dans des <e> paludismes </ e>», la sortie du module *Term Extractor* renvoie 5 termes: Antimicrobiens, Enfants, hôpital admis, paludisme, zones.
- Création d'un objet Map contenant le terme ambigu en tant que clé et la liste des termes relatifs au contexte en tant que valeur. Par exemple, la phrase précédente sera présentée par Entrée <malaria, Liste <Antimicrobiens, Enfants, hôpitaux admis, zones >>.

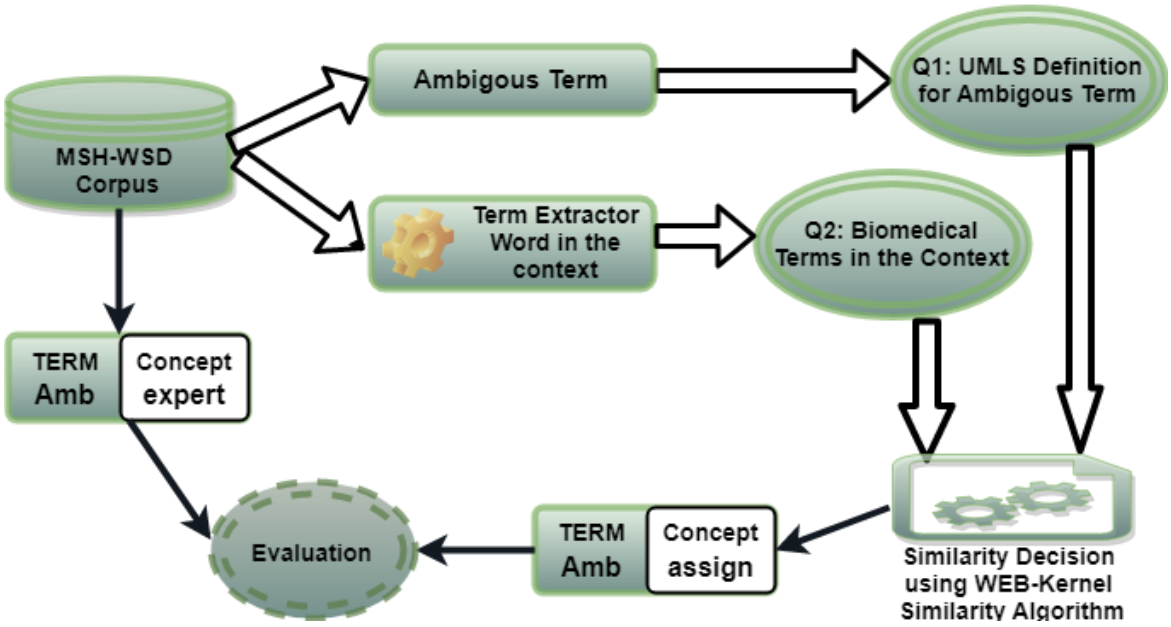


Fig 14 Organigramme du système d'évaluation - WEB-RST -

- Faire la correspondance des termes ambigus avec leurs concepts. En conséquence, la nouvelle Map contient les termes (termes ambigus et termes relatifs au contexte) avec leurs concepts : Entrée <malaria [C0206255, C0024530], Liste <Antimicrobiens, Enfants, hôpitaux admis, zones >>.
- La création d'un texte de requête pour chaque concept du terme ambigu à l'aide de leurs définitions dérivées des services de terminologie UMLS (UTS) qui fournit des services Web permettant de rechercher et d'extraire des données UMLS. Création du deuxième texte de requête en utilisant des termes dans le contexte.
- Appliquer la similarité du Web-Kernel (décrite à la sous-section 3) en utilisant les requêtes de la dernière étape, qui retournent la mesure de similarité pour chaque concept du terme ambigu, à partir des mesures retournées ; nous sélectionnons le concept approprié du terme ambigu.
- Enfin, exécuter l'étape d'évaluation pour comparer les concepts sélectionnés à l'aide de la méthode de similarité Web-Kernel et des concepts attribués par des experts du corpus MSH-WSD.

6. Résultats et discussion

Compte tenu du système présenté précédemment, nous avons effectué des expériences afin d'évaluer la méthode de similarité de Web-kernel exécutée sur le jeu de données MSH-WSD. En ce qui concerne les informations de contexte, nous utilisons une fenêtre de taille 2, ce qui signifie

deux termes entourant le terme ambigu à droite et deux termes à gauche pour déterminer le concept approprié.

Nous comparons la précision obtenue de notre méthode proposée à celle proposée par Bridget, nommée SenseRelate [65] et à celle proposée par Gad El-Rab [148].

La figure 15 présente l'exactitude de notre similarité proposée avec le web-kernel et SenseRelate en utilisant la mesure de chemin (path) et la méthode de Gad El-Rab.

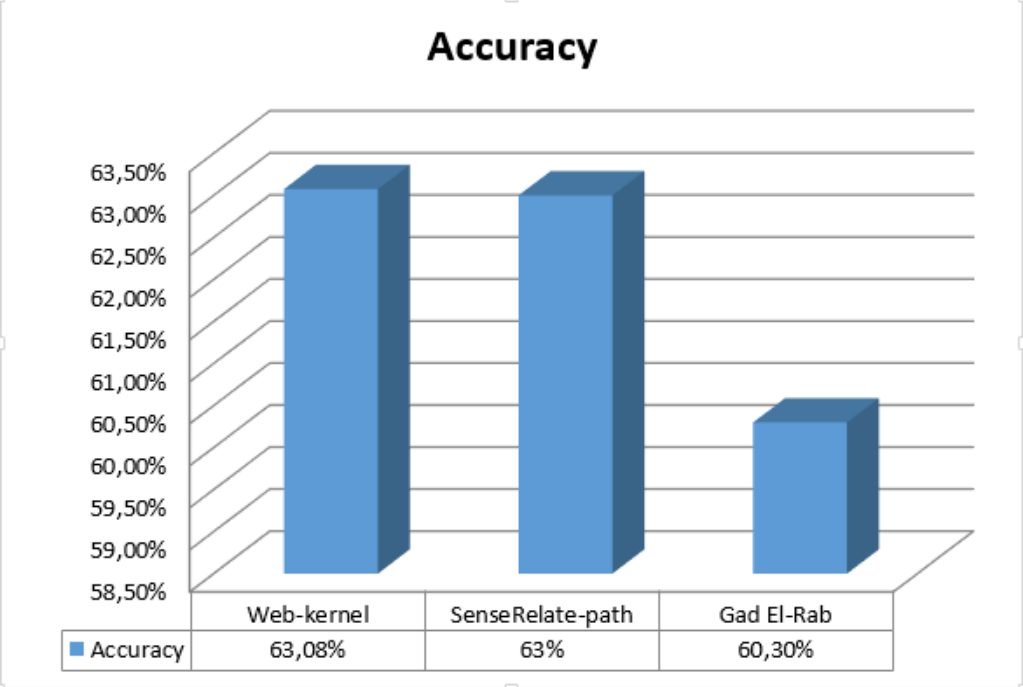


Fig 15 Résultats de la précision sur MSH-WSD

Comme illustré sur la figure 15, nous pouvons observer que l'exactitude des méthodes de similarité entre les web-kernel et de SenseRelate-path est supérieure à celle de Gad El-Rab, avec une précision de respectivement 63,03% et 63%, ce qui représente une amélioration significative de 3% sur la méthode Gad El-Rab.

Pour évaluer les performances de la méthode proposée, le tableau 7 indique les 10 précisions les plus élevées et le tableau 8 les 10 précisions les plus basses, regroupées par terme ambigu.

Tableau 7 10 Précisions Plus Grandes

Ambiguous Term	Total Example	Total Correct	Accuracy
CCD	141	139	**98,58%**
BPD	198	194	**97,98%**
BSE	198	193	**97,47%**
GAG	198	193	**97,47%**
Lawsonia	115	112	**97,39%**
IP	196	190	**96,94%**

WBS	128	123	**96,09%**
SLS	164	157	**95,73%**
INDO	122	115	**94,26%**
CLS	34	32	**94,12%**

Tableau 8 10 Précisions Les Plus Basses

Ambigous Term	Total Example	Total Correct	Accuracy
EM	129	48	**37,21%**
NM	122	44	**36,07%**
TSF	53	19	**35,85%**
Lupus	297	106	**35,69%**
Fish	198	70	**35,35%**
Sodium	197	64	**32,49%**
HGF	192	62	**32,29%**
PHA	110	27	**24,55%**
DE	126	30	**23,81%**
PCA	491	99	**20,16%**

Comme indiqué dans les tableaux 7 et 8, la méthode de similarité de Web-kernel a atteint une précision élevée de 98,58% lors de l'exécution de notre méthode sur CDD Terme ambigu du corpus MSH-WSD et une valeur de précision faible de 20,16% sur PCA Terme ambigu.

Dans ce contexte, la figure 16 présente la précision de 203 termes ambigus, et nous pouvons constater que la plupart de ces exactitudes se situent entre 45% et 98% et qui sont des valeurs intéressantes par rapport à la méthode de Gad El-Rab, la Précision la plus précise est de 97% et la faible Précision de 0%. Cela peut démontrer l'intérêt de notre méthode proposée.

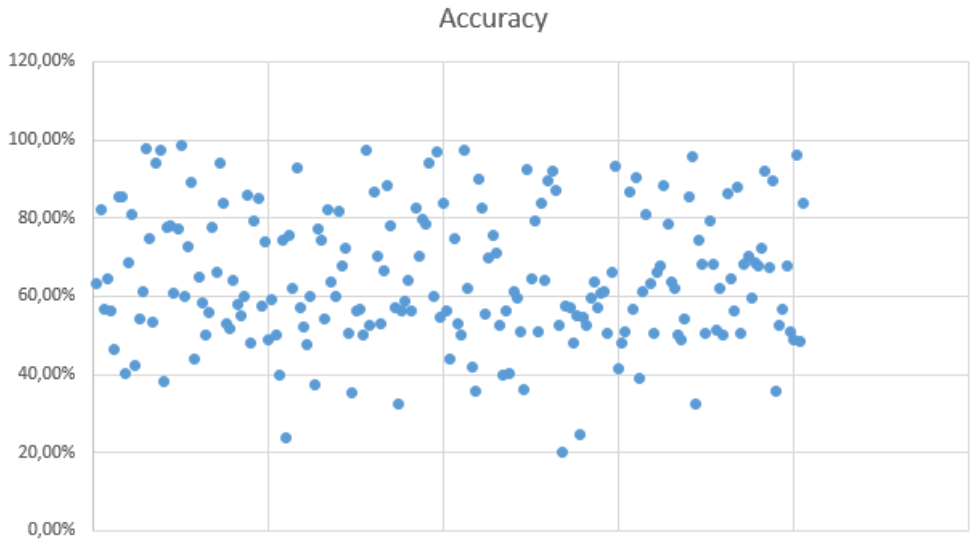


Fig 16 Précisions pour tous les termes ambigus de MSH-WSD

7. Conclusion

Les algorithmes de WSD identifient automatiquement le sens approprié d'un mot ambigu en fonction du contexte dans lequel le mot est utilisé, ces algorithmes généralement utilisent des mesures de similarité sémantique et de connexité pour quantifier la proximité sémantique entre deux concepts du contexte. Les mesures de connexité (Relatedness measures) déterminent la connexité entre deux concepts en se basant sur leurs définitions, en raison de la définition abrégée d'un concept, nous pensons qu'une similarité de texte abrégé peut être utile dans le processus du WSD pour le domaine Biomédicale.

Pour mesurer la relation sémantique entre les concepts afin de résoudre l'ambiguïté d'une expression par rapport aux multiples concepts possibles. Dans ce chapitre, nous avons proposé une nouvelle mesure de similarité au niveau du noyau Web (Web-kernel) entre une paire de textes courts biomédicaux, cette mesure utilise le grand volume de documents renvoyés par le moteur de recherche PubMed pour déterminer le contexte plus général d'un texte court biomédical par le biais d'un nouveau système de pondération des termes basé sur la théorie des ensembles approximatif.

Pour illustrer l'efficacité de notre proposition, nous avons effectué des expériences en évaluant la méthode de similarité de noyau Web exécutée sur un ensemble de données MSH-WSD. En effet, les résultats de nos expériences montrent une amélioration considérable de la précision pour les plus grandes et les plus basses, tout en regroupant les performances pour chaque terme.

Partie II : Applications - Catégorisation, Indexation/Recherche d'Information et Navigation (PubMed)

- **Chapitre 5 : Catégorisation du Texte Biomédical**
- **Chapitre 6 : Indexation et Recherche d'Information Biomédicale**
- **Chapitre 7 : Navigation et Représentations des Résultats de Recherche de PubMed**

Introduction de la partie II

Après avoir surmonté les problèmes de la Représentation du Texte Biomédical par la proposition d'une Représentation Conceptuelle, ainsi pour réaliser la première étape qui consiste à extraire l'ensemble des mots les plus significatifs, nous avons utilisé la reconnaissance des entités nommées afin de bénéficier des résultats obtenu lors de notre étude sur la reconnaissance des entités biomédical BioNER présenté dans chapitre II. La deuxième étape où ces mots sont projetés par un sens qui lui correspond le plus est due par la conceptualisation, il s'agit du processus consistant à mapper des mots littéralement présents dans du texte avec leurs concepts ou sens correspondants, en les faisant correspondre à des ressources sémantiques. Néanmoins, pour conceptualiser les documents biomédicaux, l'attribution d'un terme à ses concepts dans une ontologie peut impliquer dans certains cas le risque de perte d'informations, en raison de l'ambiguïté naturelle du thésaurus UMLS, ce qui génère de l'ambiguïté dans la représentation finale du document. Nous distinguons trois stratégies pour la désambiguïsation : tous les concepts, premier concept et basé contexte ***(All Concepts, First Concept and Based-Context)*** *(voir Chapitre III).* Nous nous sommes intéressés principalement par la troisième stratégie de désambiguïsation celle basée sur le contexte.

Pour la désambiguïsation des sens en se basant sur le contexte, afin de déterminer le sens correct (concept correct), nous utilisons généralement des mesures de similarité sémantique ou de connexité qui tentent de quantifier la proximité sémantique entre deux concepts.

Dans la partie précédente nous avons présentés deux études en relation avec les mesures de similarité et de connexité entre deux concepts par la proposition de deux méthodes ***SenseRelate/NoDistSenseRealte*** et ***Web-kernel***. Nous avons évalué ces deux méthodes sans les intégrer dans une application de fouille de texte biomédical, à travers l'utilisant du corpus MSH-WSD de NLM qui contient 203 termes ambigus et acronymes, chaque mot cible contient environ 187 exemples d'utilisation dans un résumé issue de MEDLINE 2010 et qui sont annotés par le concept adéquat définit par les experts.

Cependant, il sera de très grande importance d'évaluer la Représentation du texte biomédical proposée dans des applications de fouille de texte Biomédical, en intégrant les deux méthodes ainsi proposés.

Dans le cadre de cette partie, nous nous sommes intéressés par l'intégration de la première méthode ***SenseRelate/NoDistSenseRealte*** dans deux applications de Fouille de Texte Biomédical : La Catégorisation du Texte et l'Indexation/Recherche d'information. Pour la première application, nous essayons d'estimer l'impact des trois stratégies de désambiguïsation sur la catégorisation de texte et nous évaluons la contribution des algorithmes de WSD à la catégorisation de

texte biomédical en utilisant trois modèles d'apprentissage automatique : Machine à vecteur de support (SVM), Naïve Bayes (NB) et Entropie Maximum (ME). Pour la deuxième, nous présentons une étude pour estimer l'impact des stratégies de conceptualisation tout en appliquant les trois stratégies de désambiguïsation au processus d'indexation et nous évaluons la contribution des algorithmes de désambiguïsation de sens (WSD) au système de recherche d'informations biomédicales.

Par ailleurs, dans la même partie, nous nous sommes intéressés au problème de navigation rencontrés l'hors de l'utilisation de ***PubMed*** par les experts. En effet, nous avons proposé une troisième contribution pour surmonter le problème de navigation et permettre aux utilisateurs de ***PubMed*** un processus de navigation plus simple et efficace pour satisfaire leur besoin. Cette dernière contribution est basée sur le regroupement des résultats de recherche retourné par ***PubMed***.

Cette partie est constituée de trois chapitres :

- **Chapitre V :** Catégorisation du Texte biomédical.
- **Chapitre VI** : Indexation et Recherche d'Information.
- **Chapitre VII :** Navigation et Représentation des Résultats de Recherche de PubMed.

Chapitre 5 : Catégorisation du Texte Biomédical

1. Introduction

La Catégorisation de Texte est une application de fouille de texte dans lequel les documents sont classés en fonction de catégories prédéfinies utilisant le contenu des documents.

La gestion du nombre croissant d'articles dans les bases de données biomédicales telles que la bibliothèque numérique biomédicale MEDLINE est devenue essentielle. L'application de techniques de catégorisation de texte est essentielle pour organiser et faciliter le filtrage des articles de ces bases de données.

La catégorisation du texte biomédical passe par trois étapes essentielles : Première étape est le prétraitement et extraction des termes du lexique biomédical, la deuxième c'est le Représentation du texte biomédical sous formes de vecteurs représentants les documents, la troisième et dernière étape concerne l'apprentissage automatique de ces vecteurs et la phase de classification.

La deuxième étape concernant la représentation des documents par des vecteurs est généralement réalisée avec une représentation BoW qui souffre du manque de sens dans la représentation finale du document en raison de l'absence de toute sémantique qui réside dans le texte original, l'intégration de la sémantique est résolue par une représentation BoC. Néanmoins, en appliquant cette représentation conceptuelle, l'attribution d'un terme à ses concepts dans une ontologie peut générer de l'ambiguïté dans la représentation finale du document. Dans la partie précédente, nous avons distingué trois stratégies de désambiguïsation ***(All Concepts, First Concept and Based-Context)***. Ainsi nous avons proposé une méthode basée sur le contexte ***SenseRelate*** avec une version modifiée ***NoDistanceSenseRelate*** pour résoudre le problème de l'ambiguïté.

Dans ce chapitre, nous proposons de réaliser notre propre système de catégorisation en intégrant ***SenseRelate/NoDistanceSenseRelate*** dans le processus de représentation du texte biomédical. Notre objectif est double, d'une part nous nous sommes intéressés à réaliser une étude comparative entre les deux représentations BoW et BoC, d'autre part, de mettre en évidence l'intérêt d'utilisation du BoC pour améliorer les performances des résultats d'un tel système de Catégorisation de documents biomédicaux en réalisant une étude concernant l'impact des trois stratégies de désambiguïsation ***(All Concepts, First Concept and Based-Context)*** sur la catégorisation de texte et nous évaluons par la suite la contribution des algorithmes de WSD à la catégorisation de texte biomédical en utilisant trois modèles d'apprentissage automatique : Machine à vecteur de support (SVM), Naïve Bayes (NB) et Entropie Maximum (ME). Ces expériences sont réalisées notamment dans le domaine biomédical sur le corpus OHSUMED, en utilisant la base de connaissances spécifique au domaine biomédical UMLS.

Le reste de ce chapitre est structuré de la manière suivante : certains travaux connexes sont présentés dans la section 2. L'architecture utilisée pour notre système proposé pour la catégorisation des documents biomédicaux est présentée dans la section 3. Les résultats et la discussion de notre système d'évaluation se trouvent dans la section 4. À la fin, une conclusion qui résume notre étude du chapitre dans la section 5.

2. Travaux en connexe

La catégorisation du texte est une tâche ardue en raison de la rareté et de la grande dimension des modèles de caractéristiques des documents. Ainsi, la représentation des documents est une étape importante dans la classification des textes, qui influence principalement les résultats de la classification, en particulier dans le domaine biomédical.

La représentation des documents est généralement basée sur l'approche traditionnelle du «Sac de Mots» (Bag of Word - BoW) [149]. Néanmoins, le Sac de Mots ou simplement la représentation basée sur terme souffre du manque de sens dans la représentation finale du document en raison de l'absence de toute sémantique qui réside dans le texte original. Cependant, la représentation basée sur des concepts (Bag of Concepts - BoC) permet l'intégration de la sémantique nommée la conceptualisation qui permet l'enrichissement de la représentation des documents à l'aide de ressources de connaissances externes.

Amine et *al.* [150] ont intégré une ontologie sur le processus de classification des documents et ont conclu que la représentation basée sur des concepts fournit des meilleurs résultats. Sanchez et *al.* [153] ont présenté une méthodologie pour construire automatiquement une ontologie, en extrayant des informations du World Wide Web à partir d'un mot clé initial, en partant de l'hypothèse que la construction d'une conception ontologique robuste nécessite l'intégration de la sémantique.

Litvak et *al.* [152] ont présenté l'application d'extraction de contenu Web basée sur des ontologies pour l'analyse et la classification de documents Web dans un domaine donné. La principale contribution de ce travail consiste à utiliser une ontologie multilingue basée sur le domaine dans la représentation conceptuelle des documents. Guyot et *al.* [151] ont évalué une approche multilingue basée sur une ontologie pour la recherche d'informations multilingues. Mu-Hee et *al.* [154] ont proposé une méthode automatisée de classification de documents utilisant une ontologie, qui exprime les informations de terminologie et le vocabulaire contenus dans les documents Web au moyen d'une structure hiérarchique.

De plus, la nature complexe de la sémantique résidant dans le texte rend la classification du texte difficile en raison d'ambiguïtés difficiles à résoudre. Les difficultés sont liées au modèle de

représentation textuelle qui est le Modèle Vectoriel (***Vector Space Model - VSM***), largement utilisé, qui ne permet pas d'extraire des caractéristiques utiles et significatives à partir du texte. Considérant les classificateurs basés sur le noyau tels que les SVM, de nombreux travaux définissent la fonction du noyau sur la hiérarchie de la base de connaissances produisant des noyaux sémantiques. Séaghdha [155] a proposé plusieurs fonctions du noyau sémantique sur WordNet et prouve que les SVM fonctionnent mieux avec les noyaux sémantiques.

Des études antérieures ont proposé de nouvelles extensions du VSM traditionnel afin de dépasser ses limites. De nombreux schémas de pondération pour le modèle de VSM traditionnel sont proposés dans [156], tous visant à optimiser le poids des entités, ce qui pourrait améliorer la classification du texte. De plus, d'autres travaux démontrent quelques améliorations à l'aide de nouvelles méthodes d'extraction de fonctionnalités. Afin de surmonter la limitation de VSM liée aux mots composés. Des auteurs [157] ont proposé un modèle de Sac de Phrases (***Bag of Phrases - BoP***) au lieu du sac de mots traditionnel (BoW) qui prend fréquemment des phrases de type N-gram compte lors de l'extraction des fonctionnalités. Selon les tests effectués sur KNN, les arbres de décision, les SVM et NB, le BoP proposé surpasse le BOW d'origine. Malgré les améliorations démontrées dans ce travail, peu de limitations de VSM sont traitées.

Afin de prendre en compte les ambiguïtés et les mots polysémiques en plus des limitations précédentes, des bases de connaissances sont fréquemment utilisées. En effet, des ontologies spécifiques à un thésaurus et à un domaine peuvent aider à déterminer la sémantique exacte résidante dans le texte. Selon ces approches, le nœud de stock d'origine est transformé en BoC [158], de sorte que les vecteurs résultants sont constitués de concepts liés au texte. Ces concepts peuvent être découverts dans le texte original au moyen de connaissances de base. Il est également important d'intégrer les concepts connexes à ceux déjà adoptés lors de la phase de conceptualisation. Ce déploiement enrichit la représentation des documents et pourrait améliorer les résultats de la catégorisation.

Les auteurs [158] ont présenté un nouveau modèle de représentation utilisant des concepts extraits de la connaissance de base. L'algorithme AdaBoost est testé sur trois corpus différents afin de soutenir l'approche. Lors des expérimentations sur le corpus Reuters-21578, Wordnet est utilisé comme connaissance de base, tandis que l'ontologie MESH est utilisée avec le jeu de données OHSUMED. Différentes stratégies de désambiguïsation des sens ont été utilisées. De plus, les superconcepts de concepts spécifiques découverts dans le texte sont également intégrés dans le vecteur de concept représentant les documents texte. Ces superconcepts sont recherchés jusqu'à une distance maximale dans l'ontologie. Le déploiement de superconcept dans un modèle de représentation de document s'appelle la généralisation, ce qui permet d'améliorer les résultats, en

particulier avec les ontologies générales telles que WordNet. Néanmoins, cette stratégie ne semble pas adaptée aux bases de connaissances spécifiques à un domaine, telles que MESH.

Bai, Wang [159] ont proposé d'aligner trois ontologies générales : WordNet, OpenCyc et SUMO dans leur système et utilisent la base de connaissances résultante. Le modèle BoW traditionnel est ensuite remplacé par BoC grâce à une nouvelle indexation ontologique des documents texte au moyen de cette base de connaissances. La stratégie de contexte est utilisée afin de résoudre les ambiguïtés. La classification de texte à l'aide de SVMS est réalisée sur trois corpus : Reuters-21578, OHSUMED et 20Newsgroups. Des améliorations significatives sont démontrées en particulier avec l'ensemble de données OHSUMED. Les auteurs ont conclu que les concepts intégrés sont utiles pour des jeux de données de domaine spéciaux.

Guisse, Khelif [160] ont proposé également une approche BoC avec une généralisation utilisant un algorithme de propagation de poids afin d'attribuer des poids appropriés à des superconcepts dans l'ontologie spécifique à un domaine. Ce travail a démontré une amélioration significative de la classification des brevets.

De nouveaux modèles de représentation utilisant des parties de la hiérarchie des ontologies sont également proposés dans la littérature. Ces parties constituent des arbres sémantiques [161] dans lesquels chaque concept se voit attribuer un score d'importance. En utilisant la hiérarchie sémantique de WordNet, une amélioration significative est démontrée lors de la classification de document Yahoo! [161]. Les forêts de concepts [162], constituées de parties de WordNet, contribuent à l'amélioration des résultats de la classification lorsqu'ils sont testés sur Reuters-21578.

Selon la littérature, les auteurs semblent ne pas être d'accord pour évaluer l'importance de l'utilisation de l'information sémantique dans la classification [163]. Néanmoins, il semble que cette approche soit prometteuse compte tenu du contexte particulier de la tâche de classification [100].

Dans cette étude, nous utilisons le thésaurus UMLS en tant que ressources spécifiques au domaine afin de réaliser le processus de conceptualisation.

3. Système de Catégorisation du Texte Biomédical

Afin de comparer les stratégies de désambiguïsation décrites dans le chapitre III, nous avons développé un système de catégorisation de texte biomédical pour évaluer la contribution des méthodes de désambiguïsation des sens des mots à l'aide de trois algorithmes d'apprentissage automatique : machine à vecteurs de support (SVM), Naïve Bayes (NB) et Maximum Entropy (ME). Dans cette section, nous présentons les algorithmes d'apprentissage automatique, puis nous présentons un ensemble de données et un organigramme de notre système d'évaluation.

3.1. Algorithmes d'apprentissage automatique

Une fois que le processus de conceptualisation est terminé et que la désambiguïsation est effectuée, l'étape finale de la catégorisation de texte consiste en une classification de document. Comme cela se fait toujours dans l'apprentissage automatique, pour classer un document, nous construisons des modèles d'apprentissage automatique et nous exposons les vecteurs de désarchivage du document à classer sur le modèle crée lors de la phase d'apprentissage, afin de déterminer la classe ou le niveau adéquat de catégorie du nouveau document. Bien entendu, le document à classer est traité de la même manière que les autres documents de la phase d'apprentissage. Dans ce chapitre, nous proposons d'utiliser des modèles d'apprentissage automatique : Machine à vecteur de support (***SVM***), naïve bayésien (***NB***) et entropie maximal (***ME***) sur notre système d'évaluation. La section suivante présente le corpus utilisé dans notre système d'évaluation.

3.2. Corpus

Ohsumed corpus [164], est un sous-ensemble de la base de données MEDLINE, qui est une base de données bibliographique d'importantes publications médicales revues par des pairs, gérées par la bibliothèque nationale de médecine (National Library of Medicine - NLM). Nous examinons 20 000 documents sur les 50 216 abrégés médicaux de l'année 1991, regroupés dans 23 catégories de sujets médicaux (MeSH) du groupe des maladies cardiovasculaires. Parmi les 23 catégories du groupe des maladies cardiovasculaires, nous considérons 10 catégories pour notre système d'évaluation : Infections bactériennes et mycoses, Maladies virales, Maladies parasitaires, Néoplasmes, Maladies de l'appareil locomoteur, Maladies de l'appareil digestif, Maladies stomatognathiques, Maladies des voies respiratoires, Maladies oto-rhino-laryngologiques, maladies du système nerveux (Bacterial Infections and Mycoses, Virus Diseases, Parasitic Diseases, Neoplasms, Musculoskeletal Diseases, Digestive System Diseases, Stomatognathic Diseases, Respiratory Tract Diseases, Otorhinolaryngologic Diseases, Nervous System Diseases).

3.3. Organigramme du Système d'evaluation

Comme mentionné ci-dessus, nous utilisons Ohsumed [164] comme jeu de données pour évaluer la contribution des stratégies de désambigüisions à la catégorisation du texte biomédical. La figure 17 présente l'organigramme du système d'évaluation.

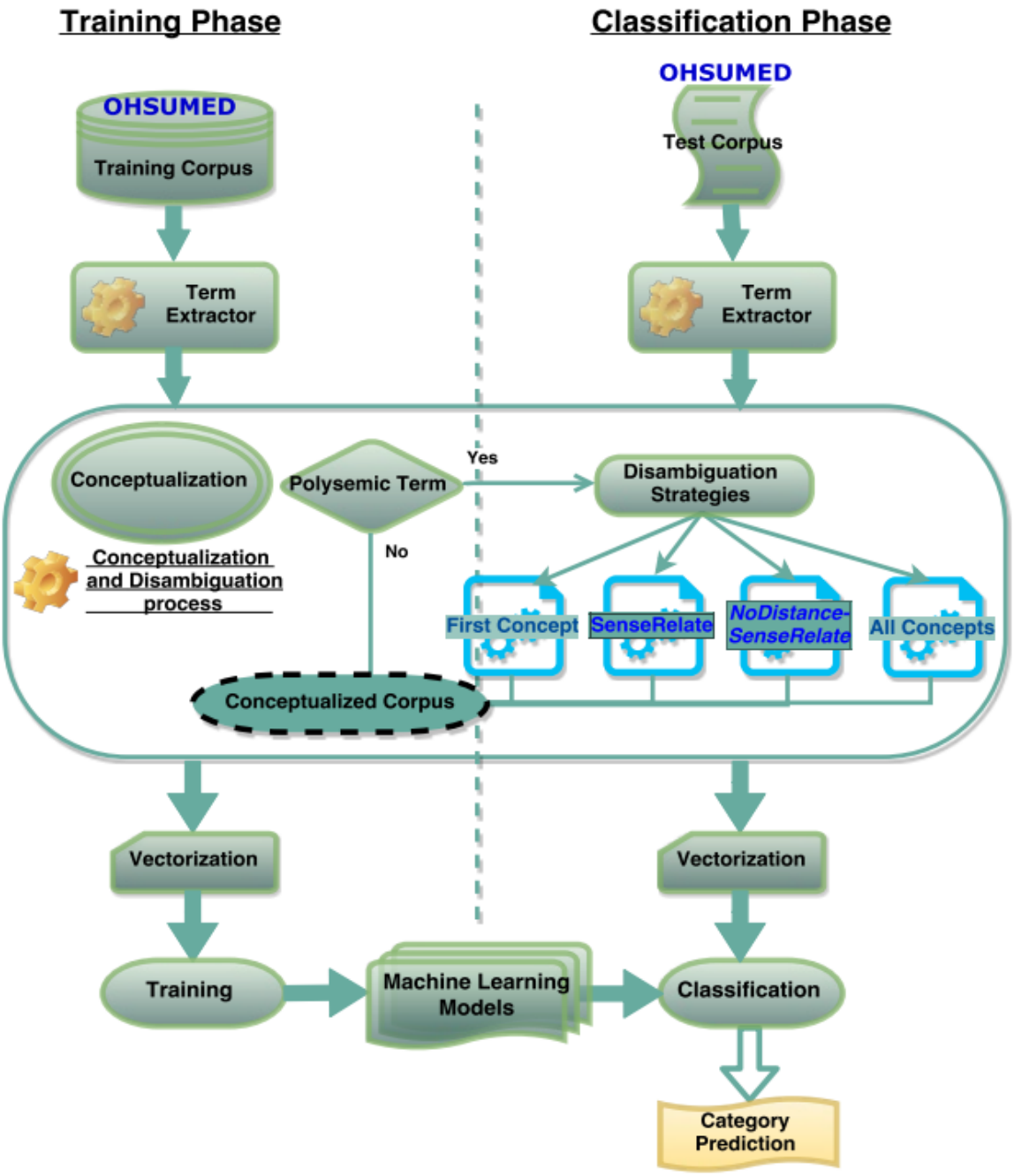


Fig 17 Système proposé pour la catégorisation de documents biomédicaux

Dans ce système, nous proposons de procéder comme suit :

a. Extraction des Termes

Exécution du module d'extraction de termes pour le domaine biomédical qui tente d'extraire des termes biomédicaux du corpus en utilisant UMLS comme dictionnaire.

b. Conceptualisation

La projection de tous les termes sur les concepts correspondants à l'aide de l'UTS (UMLS Terminology Services) qui fournit des services Web permettant de rechercher et d'extraire des données UMLS.

c. Désambiguïsation

Le processus de désambiguïsation est lancé pour les termes polysémiques, nous utilisons les stratégies de désambiguïsation suivantes : premier concept, tous les concepts, basé sur le contexte (SenseRelate, NoDistanceSenseRelate). Donc, pour chaque stratégie, le document sera représenté par un Sac de Concepts.

d. Apprentissage Automatique

Faire apprendre les Vecteurs des documents à l'aide de modèles d'apprentissage automatique.

e. Classification

Utilisation de ces modèles de classificateurs pour classer les vecteurs de documents de test.

4. Résultats et Discussion

Compte tenu du système présenté précédemment, nous avons effectué des expériences afin d'évaluer l'effet des stratégies de désambiguïsation sur la catégorisation de texte biomédical à l'aide de trois algorithmes d'apprentissage automatique : NB, SVM et ME.

Dans cette section, nous présentons les mesures de performance, les résultats de classification obtenus et enfin, nous analysons et discutons les résultats obtenus.

4.1. Résultats

Dans nos expériences, nous avons testé toutes les méthodes sur le corpus OHSUMED après la conceptualisation réalisée selon quatre stratégies de désambiguïsations différentes.

Trois algorithmes d'apprentissage automatique ont été utilisés dans nos expériences : SVM, NB et ME.

Nous avons évalué les assignations des catégories du système de classification en utilisant la précision, rappel et F1-mesure [165]. Pour calculer ces métriques, nous avons d'abord calculé les valeurs des paramètres suivants pour chaque classificateur :

Vrai Positif (True Positive – TP) : fait référence à l'ensemble des documents qui sont correctement assignés à une catégorie donnée.

Faux Positif (False Positive – FP) : fait référence à l'ensemble de documents mal assignés à une catégorie.

Faux Négatif (False Negative - FN) : fait référence à l'ensemble de documents qui sont incorrectes et qui ne sont pas affectés à la catégorie.

Ensuite, nous avons calculé la précision et le rappel pour chacun des classificateurs avec les formules suivantes :

$$P = \frac{TP}{TP+FP} \quad (5.1)$$

$$R = \frac{TP}{TP+FN} \quad (5.2)$$

$$F1\ measure = 2\frac{P*R}{P+R} \quad (5.3)$$

4.1.Discussion

Le tableau 9 montre la mesure F1 obtenue en utilisant ME, NB et SVM pour la catégorisation des documents OHSUMED. Comme mentionné dans la section (3.3), nous utilisons dix catégories du corpus OHSUMED pour le système de catégorisation de texte en utilisant d'abord les trois méthodes de classification sur le corpus original OHSUMED sans conceptualisation (basée sur le terme). Les méthodes sont ensuite testées à l'aide de chacune des trois stratégies de désambiguïsation: Tous les concepts, Premier concept, basé sur le contexte (SenseRelate et NoDistanceSenseRelate).

Tableau 9 Résultats de la désambiguïsation Stratégies utilisant des modèles d'apprentissage automatique

	Bacterial Infections and Mycoses		**Virus Diseases**		**Parasitic Diseases**		**Neoplasms**		**Musculoskeletal Diseases**	
	F1-Measure	***Baseline Gap***	***F1-Measure***	***Baseline Gap***	***F1-Measure***	***Baseline Gap***	***F1-Measure***	***Baseline Gap***	***F1-Measure***	***Baseline Gap***
	Classificateur Entropie Maximal									
Term-Based (Baseline)	***72,90%***		***77,60%***		***87,71%***		***70,94%***		***74,35%***	
First Concept	72.08%	**-0,83%**	76.85%	**-0,75%**	70.27%	**+0,58%**	70,27%	**-0,67%**	70,27%	**-0,35%**
All Concepts	71.18%	**-1,72%**	75.31%	**-2,28%**	70.77%	**-1,08%**	70,77%	**-0,16%**	70,77%	**-1,70%**
SenseRelate	76.53%	**+3,63%**	81.26%	**+3,66%**	74.85%	**+3,24%**	74,85%	**+3,91%**	74,85%	**+4,08%**
NoDistance-SenseRelate	76.55%	**+3,64%**	81.40%	**+3,80%**	74.65%	**+2,97%**	74,65%	**+3,71%**	74,65%	**+4,01%**
	Classificateur Naïve Bayésien									
Term-Based (Baseline)	***72,14%***		***76,91%***		***87,76%***		***69,66%***		***76,07%***	
First Concept	71.35%	**-0,78%**	77.14%	**+0,23%**	88.08%	**+0,32%**	69,45%	**-0,21%**	76,24%	**+0,16%**
All Concepts	69.75%	**-2,38%**	76.32%	**-0,58%**	86.69%	**-1,07%**	69,63%	**-0,03%**	75,59%	**-0,48%**
SenseRelate	75.74%	**+3,60%**	81.72%	**+4,81%**	91.03%	**+3,27%**	74,44%	**+4,77%**	80,84%	**+4,76%**
NoDistance-SenseRelate	75.78%	**+3,65%**	81.74%	**+4,83%**	90.82%	**+3,06%**	74,17%	**+4,51%**	80,39%	**+4,32%**
	Classificateur Support à Vaste Marge									
Term-Based (Baseline)	***74,17%***		***78,43%***		***88,98%***		***71,13%***		***78,24%***	
First Concept	73.68%	**-0,49%**	78.36%	**-0,07%**	89.09%	**+0,11%**	71,79%	**+0,66%**	76,78%	**-1,46%**
All Concepts	72.42%	**-1,75%**	77.83%	**-0,60%**	87.46%	**-1,52%**	72,53%	**+1,40%**	76,14%	**-2,10%**
SenseRelate	78.31%	**+4,14%**	80.88%	**+2,45%**	92.41%	**+3,43%**	76,44%	**+5,31%**	81,35%	**+3,11%**

NoDistance-SenseRelate	*78.36%*	***+4,19%***	*80.94%*	***+2,51%***	*92.22%*	***+3,24%***	*74,49%*	***+3,36%***	*81,27%*	***+3,03%***
	Digestive System Diseases		**Stomatognathic Diseases**		**Respiratory Tract Diseases**		**Otorhinolaryngo-logic Diseases**		**Nervous System Diseases**	
	Classificateur Entropie Maximal									
Term-Based (Baseline)	***74,63%***		***74,66%***		***68,18%***		***77,96%***		***69,44%***	
First Con-cept	*73,55%*	***-1,08%***	*76,24%*	***+1,58%***	*69,82%*	***+1,64 %***	*78,17%*	***+0,21%***	*69,27%*	***-0,16%***
All Concepts	*72,74%*	***-1,89%***	*75,81%*	***+1,15%***	*69,03%*	***+0,85 %***	*77,09%*	***-0,88%***	*69,76%*	***+0,32%***
SenseRelate	*77,88%*	***+3,25%***	*80,82%*	***+6,16%***	*74,24%*	***+6,07 %***	*82,47%*	***+4,51%***	*73,92%*	***+4,48%***
NoDistance-SenseRelate	*78,18%*	***+3,54%***	*80,75%*	***+6,09%***	*74,42%*	***+6,25 %***	*82,43%*	***+4,47%***	*73,83%*	***+4,39%***
	Classificateur Naïve Bayésien									
Term-Based (Baseline)	***76,16%***		***80,12%***		***71,28%***		***78,21%***		***71,99%***	
First Con-cept	*76,14%*	***-0,02%***	*80,00%*	***-0,13%***	*71,18%*	***-0,09%***	*77,48%*	***-0,73%***	*72,04%*	***+0,05%***
All Concepts	*75,45%*	***-0,71%***	*79,60%*	***-0,52%***	*70,47%*	***-0,81%***	*77,52%*	***-0,69%***	*71,31%*	***-0,68%***
SenseRelate	*80,47%*	***+4,32%***	*84,40%*	***+4,28%***	*75,54%*	***+4,26%***	*81,83%*	***+3,62%***	*76,70%*	***+4,71%***
NoDistance-SenseRelate	*80,69%*	***+4,53%***	*84,24%*	***+4,12%***	*75,59%*	***+4,32%***	*81,91%*	***+3,69%***	*76,32%*	***+4,33%***
	Classificateur Support à Vaste Marge									
Term-Based (Baseline)	***78,76%***		***80,93%***		***70,50%***		***81,59%***		***71,43%***	
First Con-cept	*77,26%*	***-1,50%***	*81,91%*	***+0,98%***	*71,55%*	***+1,06%***	*80,65%*	***-0,95%***	*71,26%*	***-0,17%***
All Concepts	*77,21%*	***-1,55%***	*80,99%*	***+0,06%***	*70,79%*	***+0,29%***	*79,39%*	***-2,21%***	*71,41%*	***-0,01%***
SenseRelate	*81,96%*	***+3,20%***	*83,60%*	***+2,67%***	*76,63%*	***+6,13%***	*84,98%*	***+3,39%***	*76,06%*	***+4,63%***
NoDistance-SenseRelate	*82,14%*	***+3,38%***	*83,29%*	***+2,36%***	*76,16%*	***+5,66%***	*85,16%*	***+3,56%***	*75,90%*	***+4,47%***

Comme illustré dans le tableau 9, nous pouvons observer dans tous les cas, la désambiguïsation en utilisant SenseRelate et NoDistanceSenseRelate améliore le résultat.

Considérant chaque catégorie indépendamment dans les résultats, nous pouvons observer que NoDistanceSenseRelate surpasse les autres stratégies dans environ 94% des cas pour les catégories «Infections et mycoses bactériennes», «Maladies virales», «Maladies de l'appareil digestif» et «Maladies de l'appareil respiratoire», en utilisant les modèles NB, ME et SVM. SenseRelate surpasse les autres stratégies d'environ 93% pour les catégories «Maladies parasitaires», «Tumeurs», «Maladies musculo-squelettiques», «Maladies stoma-tognathiques» et «Maladies oto-rhino-laryngologiques» à l'aide de modèles NB, ME et SVM.

La stratégie NoDistanceSenseRelate surpasse la représentation traditionnelle (basée sur le terme) ainsi que First Concept and All Concepts, atteignant une mesure F1 de 81,40% sur la catégorie «Virus Diseases», elle présente une amélioration significative de 3,80% par rapport à la représentation basée sur le terme, 4,47% sur la stratégie du premier concept et 5,36% sur la stratégie de tous les concepts lors de l'utilisation de ME.

La stratégie SenseRelate surpasse celle basée sur terme, First concept et All concepts, atteignant une mesure F1 de 76,70% sur la catégorie «Maladies du système nerveux». Elle présente une amélioration significative de 4,66% par rapport à la stratégie du First concept, 4, 71% sur la représentation basée sur la durée et 5,39% sur la stratégie All concepts lorsque nous utilisons NB.

Les résultats de la représentation basée sur les termes en général sont très proches de la performance du First concept, tandis que les deux stratégies surperforment la stratégie de All concepts dans environ 70% des cas pour toutes les catégories en raison de l'augmentation de l'espace de représentation.

Les résultats globaux pour toutes les stratégies de désambiguïsation montrent que la classification utilisant le modèle SVM obtient un score plus élevé, atteignant une mesure F1 de 81,96% sur «Maladies de l'appareil digestif» à l'aide de SenseRelate, ce qui représente une amélioration de 1,49% par rapport au deuxième meilleur modèle. NB, comparé à ME, NB présente une amélioration de 2,59%.

5. Conclusion

L'ambiguïté des sens des mots (WSD) joue un rôle essentiel dans de nombreuses applications d'extraction de texte biomédical telles que la catégorisation. ***SenseRelate*** et ***NoDistanceSenseRelate*** sont des algorithmes de WSD basés sur le contexte.

Pour illustrer l'efficacité de ces algorithmes, ***SenseRelate*** et ***NoDistanceSenseRelate*** ont été intégrés dans notre système de catégorisation de texte biomédical. Plusieurs expériences ont été menées à l'aide de trois modèles d'apprentissage automatique: SVM, NB et ME.

Les résultats obtenus avec le corpus OHSUMED montrent que les méthodes basées sur le contexte (***SenseRelate*** et ***NoDistanceSenseRelate***) sont plus performantes que les autres.

Chapitre 6 : Indexation et Recherche d'Information Biomédicale

1. Introduction

L'indexation des documents consiste à extraire, organiser, structurer, indexer et sauvegarder le contenu sémantique des documents dans des structures de données appelées index. L'indexation permet de retrouver rapidement les documents contenant un ou plusieurs termes donnés et éventuellement ses positions, ses fréquences d'occurrence dans un document. Nous distinguons deux types d'indexation : Indexation libre et Indexation contrôlée. Dans l'indexation libre, l'indexeur extrait les mots-clés d'un document ou les choisit librement sans l'aide de sources de connaissance. Alors que dans l'indexation contrôlée : le vocabulaire ou le langage d'indexation est déjà défini a priori et son utilisation exclusive s'impose à l'indexeur. Notons que ces deux types d'indexation peuvent se réaliser manuellement ou automatiquement. L'indexation contrôlée automatique ou indexation sémantique est plus robuste vue qu'elle se base sur des ressources de connaissances terminologiques.

La représentation du document est une étape essentielle et critique dans le processus d'indexation de document, afin de représenter un document en générant le modèle d'indexation, deux solutions de représentations possible BoW et BoC. Comme nous avons étudié dans la partie 1, la représentation BoW néglige souvent le sens dans la représentation finale du document en raison de la non utilisation de ressources sémantiques telles que des dictionnaires ou des ontologies, cependant, la conceptualisation dans la représentation BoC utilise des ressources de connaissances qui permet l'enrichissement des modèles de représentation de document, trois stratégies peuvent être utilisées pour réaliser la tâche de conceptualisation: ajout du concept (*Adding Concept*), conceptualisation partielle (Partiel Conceptualization) et conceptualisation complète (Complet Conceptualization).

Néanmoins, tout en réalisant la tâche de conceptualisation en mappant terme à concept à l'aide d'une ontologie, plusieurs correspondances peuvent être déduites en raison de l'ambiguïté naturelle du thésaurus du système de langage médical unifié (UMLS), puis obtenir un document ambigu causer par cette tâche de mappage, pour surmontant ce problème, nous avons distingué trois stratégies de désambiguïsation : Tous les concepts (All concept), Premier Concept (First concept) et contexte basé (Context Based).

Au cours des chapitres précédents, nous avons évalué les algorithmes de WSD en utilisant plusieurs stratégies de désambiguïsation ainsi que des mesures sémantiques et de connexité. Dans ce chapitre, nous proposons d'étudier l'impact des stratégies de conceptualisation tout en appli-

quant les trois stratégies de désambiguïsation au système d'indexation de document et nous évaluons la contribution des algorithmes de WSD au système de Recherche d'Informations Biomédicales.

Le reste de ce chapitre est structuré selon la séquence suivante : certains travaux connexes sont présentés dans la section 2. La section 3 est consacrée pour le système de recherche d'information et l'architecture proposé pour notre système d'indexation et de recherche d'information. Les résultats et discussion de notre système proposé se trouvent dans la section 4. Nous concluons ce chapitre dans la section 5.

2. Travaux en connexe

De nombreux travaux dans la littérature ont contribué dans le contexte étudié de ce chapitre, Amine et *al.* [150] ont proposé d'intégrer une ontologie dans le processus de classification de documents et ont mis en évidence que la représentation basée sur des concepts fournit les meilleurs résultats. Litvak et *al.* [152] ont introduit l'application d'extraction de contenu Web basée sur une ontologie pour l'analyse et la classification de documents Web dans un domaine donné ; La principale contribution de ce travail consiste à utiliser une ontologie multilingue basée sur le domaine dans la représentation conceptuelle des documents. Guyot et *al.* [151] ont évalué une approche multilingue basée sur des ontologies pour la recherche d'informations multilingues. Mu-Hee et *al.* [154] ont suggéré une méthode automatisée de classification de documents utilisant une ontologie, qui exprime des informations terminologiques et le contenu vocabulaire dans des documents Web à travers une structure hiérarchique.

Par ailleurs, l'utilisation de la représentation conceptuelle peut impliquer dans certains cas le risque de perte d'informations. Ensuite, le choix du sens le plus approprié pour un terme en appliquant des algorithmes de WSD aura une influence négative sur les performances de l'IR.

Sanderson [166] a présenté une étude des travaux sur la désambiguïsation des sens des mots et la recherche d'informations et il a examiné les approches permettant d'améliorer l'efficacité de la récupération faisant appel à la connaissance des sens.

D'autres études ont utilisé les sens sélectionnés en désambiguïsant les termes dans les requêtes et les documents pour effectuer l'indexation et la recherche. Stokoe et *al.* [167] ont étudié l'utilisation d'un système WSD automatisé à la pointe de la technologie dans un cadre Web de IR. L'objectif de leur recherche est de réaliser une évaluation à grande échelle de l'algorithme WSD automatisé et du système IR. Leur objectif est de démontrer les performances relatives d'un système IR utilisant WSD par rapport à une technique d'extraction de base telle que le modèle d'espace vectoriel.

Kim et *al.* [168] ont proposé une méthode de marquage sensoriel souple, cohérente pour améliorer les performances de recherche de texte à grande échelle. En effet, leur étiqueteur de sens peut être construit sans corpus avec étiquette de sens et effectue une homonymie cohérente en considérant uniquement le mot voisin le plus informatif comme preuve de la détermination du sens du mot cible.

D'autres chercheurs ont proposé d'utiliser les sources de connaissances depuis des thésaurus pour étudier l'effet d'expansion. Le but de l'article de Fang [169] est d'étudier si l'expansion de requêtes utilisant uniquement des ressources lexicales créées manuellement pourrait conduire à une amélioration des performances. La principale contribution de son travail est de montrer que l'expansion des requêtes en utilisant uniquement des ressources lexicales conçues à la main est efficace dans le cadre axiomatique récemment proposé. Agirre et *al.* [170] ont proposé d'utiliser WordNet pour l'expansion de documents en proposant une nouvelle méthode : à partir d'un document complet, un algorithme de parcours aléatoire du graphe WordNet classe les concepts étroitement liés aux mots du document.

Dans le domaine biomédical, certaines études [172] [171] ont étudié l'utilisation du thésaurus MeSH pour créer un modèle d'indexation permettant d'améliorer les résultats du processus de recherche. Dinh et *al.* [172] ont proposé d'évaluer une approche basée sur le sens pour indexer et rechercher des documents biomédicaux à l'aide de deux méthodes WSD permettant d'identifier des concepts MeSH ambigus : de gauche à droite et sur un cluster, puis ils ont intégré ces deux méthodes dans un cadre d'indexation et de recherche de documents basé sur le sens. Majdoubi et *al.* [171], ont proposé une indexation conceptuelle des articles médicaux à l'aide du thésaurus MeSH (vedettes-matière médicales), appelé BIOINSY (système d'indexation BIOmédical), et utilisent l'approche du modèle linguistique pour comprendre le sens du terme et en déterminer le descripteur dans le contexte du document.

3. Système d'Indexation et Recherche d'Information Biomédicale

3.1. Indexation des Documents Biomédicaux

L'indexation des documents consiste à extraire, organiser, structurer, indexer et sauvegarder le contenu sémantique des documents dans des structures de données appelées index. L'indexation permet de retrouver rapidement les documents contenant un ou plusieurs termes donnés et éventuellement ses positions, ses fréquences d'occurrence, ...etc. dans un document.

A. Type d'indexation

Nous distinguons deux types d'indexation : Indexation libre et Indexation contrôlée

a. Indexation libre :

L'indexeur extrait les mots-clés d'un document ou les choisit librement sans l'aide de sources de connaissance. L'indexation peut être améliorée en filtrant les mots fonctionnels pour éliminer les non-descripteurs du document. L'avantage de cette indexation est que le traitement est plus facile et rapide en amont, mais le résultat présent de très grandes difficultés à l'interrogation : l'atomisation du vocabulaire (l'ensemble de descripteurs le constituant n'est pas connu a priori) ou l'ambiguïté des termes choisis par rapport aux sujets du document, ...

b. Indexation contrôlée :

Le vocabulaire ou le langage d'indexation est déjà défini a priori et son utilisation exclusive s'impose à l'indexeur. Ce vocabulaire peut se présenter sous deux formes : (1) liste alphabétique des descripteurs simples et (2) liste structurée de descripteurs reliés entre eux par des relations de hiérarchie, d'association ou d'équivalence. Les descripteurs choisis dans l'index ne sont pas nécessairement observables directement dans les documents à indexer. Dans le cas où il s'agit d'un domaine spécifique, le vocabulaire est appelé terminologie du domaine ou aussi terminologie de référence.

Notons que ces deux types d'indexation peuvent se réaliser manuellement ou automatiquement. L'indexation manuelle est une opération réalisée par un spécialiste ou un documentaliste qui consiste à recenser les sujets ou concepts dont traite un document et à les représenter à l'aide d'un vocabulaire contrôlé prédéfini. L'indexation automatique se base sur des méthodes statistiques et probabilistes permettant de repérer des éléments significatifs dans chaque document. Bien que les vocabulaires contrôlés soient souvent utilisés pour l'indexation manuelle, plusieurs travaux de recherche ont été abordés dans la littérature pour une indexation contrôlée automatique en utilisant des ressources terminologiques nommée Indexation Sémantique.

B. Indexation Sémantique

L'indexation sémantique consiste, lors de l'analyse d'un document à rattacher chaque mot à un concept sous-jacent qui représente le sens des mots. Ainsi, par exemple, pour le mot "jaguar", il faut déterminer s'il s'agit du félin, de la voiture ou de l'avion. Dans le travail de Baziz [179], l'indexation sémantique a été définie comme une approche d'indexation basée sur le sens des mots (appelée sense-based indexing en anglais). Plusieurs travaux d'indexation sémantique ont été menés dans ce sens [180] [181] [182] [183] [184] [179] [185].

De manière générale, le principe de base de l'indexation sémantique consiste à extraire dans un premier temps, l'ensemble des mots les plus significatifs. Ces mots sont par la suite affectés

par un sens qui lui correspond le plus. Si un mot-clé possède plusieurs sens, celui-ci est désambiguïsé pour retenir le sens le plus adéquat dans son contexte d'utilisation.

Dans le cadre de ce chapitre et pour effectuer ce processus d'indexation dans le domaine biomédical, nous avons intégré la sémantique avec une représentation conceptuelle, afin d'étudier l'impact des stratégies de conceptualisation (***Adding Concept***, ***Concept Only***) en appliquant les trois stratégies de désambiguïsation décrites dans le chapitre 3 (***All Concept, First Concept, Based Context***), au système de recherche d'informations, nous avons développé un système d'indexation de texte biomédical basé sur les stratégies de désambiguïsation. Pour réaliser la deuxième étape du système de recherche d'informations, nous avons intégré le modèle d'indexation dans le système de recherche d'informations de Terrier [186]. Dans cette section, nous présentons le système de recherche d'informations, puis nous présentons le corpus utilisé et l'organigramme de notre système d'évaluation.

3.2. Processus de Recherche d'Informations

L'objectif principal d'un système de recherche d'informations est d'obtenir les documents les plus pertinents pour répondre à une requête de besoin d'informations à partir d'un ensemble de ressources des documents spécifiques. Pour réaliser ce processus, notre approche proposée consiste à extraire les concepts du métathésaurus pertinents depuis les documents et les requêtes à l'aide d'un thésaurus sémantique biomédical, puis représenter les documents et les requêtes par un ensemble de concepts déduits du métathesaurus à l'aide de deux stratégies de conceptualisation produisant deux vecteurs descripteurs pour chaque document/requête. Chaque concept de métathésaurus ambigu (Polysemic Term) est étiqueté avec le sens le plus probable, compte tenu de son contexte basé sur les relations sémantiques entre lui et ses mots voisins dans une phrase. Formulant le modèle d'indexation en fonction des vecteurs descripteurs des documents, le processus de recherche calcule un score numérique indiquant le degré de pertinente de chaque document par rapport à une requête donnée et classe les documents en fonction de ce score. Deux étapes sont essentielles dans notre système d'indexation et de recherche basé sur le sens.

A. Création du vecteur de document et de requête.

Étant donné le texte du document/requête, la recherche des concepts de métathésaurus pertinents formule le vecteur descripteur initial du document V^d/requêtes V^q contenant le terme biomédical annoté T_j marqué par leurs concepts candidats :

$$V_i^d = \{T_{1i}^d, T_{2i}^d, \ldots T_{ni}^d\} \quad (6.1)$$

$$V_i^q = \left\{T_{1i}^q, T_{2i}^q, \cdots T_{mi}^q\right\} \quad (6.2)$$

Où V_i^d , V_i^q sont respectivement l'ensemble des concepts du métathesarus, n et m sont respectivement les numéros de terme dans le document et la requête,T_{ji}^d , T_{ji}^q sont respectivement le j[ème] concept de métathésaurus dans le document di et la requête qi. Formellement:

$$T_j = \left\{C_{1j}, C_{2j}, \cdots C_{pj}\right\} \quad (6.3)$$

Où Ckj est le k-ième concept candidat pour le terme Tj.

Par conséquent, les vecteurs descripteurs pour les documents et les requêtes seront représentés par un ensemble de concepts. Pour le terme polysémique (Ti avec plusieurs concepts), nous utilisons les stratégies de désambiguïsation. Enfin la construction du modèle d'indexation en utilisant les vecteurs de sens obtenus VS_i^d et VS_i^q. Formellement :

$$VS_i^d = \left\{C_{1i}^d, C_{2i}^d, \cdots C_{ni}^d\right\} \quad (6.4)$$

$$VS_i^q = \left\{C_{1i}^q, C_{2i}^q, \cdots C_{mi}^q\right\} \quad (6.5)$$

Où C_{ji}^d, C_{ji}^d sont respectivement le j[ème] concept déduit de ,T_{ji}^d, T_{ji}^q dans le document di et la requête q_i.

B. Classement des documents les plus pertinents

Le document est classé par ordre décroissant de pertinence prédite par rapport à la requête en calculant la valeur du score de pertinence du document par rapport à la requête à l'aide de modèles de pondération de document.

Dans notre cas, nous avons choisi d'expérimenter par deux modèles de pondération : les modèles TF-IDF et Okapi BM25.

a. TF-IDF.

TF-IDF (Term Frequency inverce Document Frequency) [177] attribue un poids élevé à un terme, s'il apparaît fréquemment dans le document mais rarement dans l'ensemble de la collection de documents [176]. Le poids TF-IDF [177] est composé de deux termes : le premier, calcule la fréquence de terminaison normalisée (TF), autrement dit : Le nombre de fois qu'un mot apparaît dans un document N^w, divisé par le nombre total des mots dans ce document N^{wd};

$$TF(t) = N^w \Big/ N^{wd} \quad (6.6)$$

$$IDF(t) = \log\_ e\left(\frac{N^{d}}{N_{t}}\right) \tag{6.7}$$

$$Score(d, Q) = \sum[tf * \log 2(IDF)] \tag{6.8}$$

Le second terme est la fréquence de document inverse (IDF), calculée comme étant le logarithme du nombre des documents du corpus N^d divisé par le nombre des documents contenant le terme spécifique.

b. Okapi BM25.

Okapi BM25 est l'une des fonctions de pondération «simples» les plus puissantes et s'est révélée être une base de référence utile pour les expériences et les fonctions de classement.

$$\sum_{i:d_i=q_i=1} \left[\log\left(\frac{(r_i+0.5)/(R-r_i+0.5)}{(df_i-r_i+0.5)/(D-R-df_i+r_i+0.5)}\right) \cdot \frac{tf_{i,d}+k_1\cdot tf_{i,d}}{tf_{i,d}+k_1((1-b)+b\cdot\frac{dl}{avg(dl)})} \cdot \frac{tf_{i,q}+k_2\cdot tf_{i,q}}{tf_{i,q}+k_2}\right] \tag{6.9}$$

- Le score de classement de type IDF comme défini dans la section précédente ;
- La fréquence du terme de document $t_{j,d}$ normalisée par le rapport de la longueur du document dl à la longueur moyenne Avg (dl) ;
- La fréquence du terme de la requête $t_{i,q}$.

3.3. Corpus

Pour évaluer la contribution des stratégies de désambiguïsation au système d'indexation et de recherche d'informations biomédicales, nous utilisons la collection de tests Ohsumed [164] telle qu'elle a été utilisée pour TREC-9 Filtering Track, qui est un ensemble de 348 566 références provenant de MEDLINE, la base de données en ligne d'informations médicales, de titres et / ou de résumés de 270 revues médicales sur une période de cinq ans (1987-1991).

Les champs disponibles sont :

1. Le titre ;
2. Le résumé ;
3. Les termes d'indexation MeSH, l'auteur ;
4. La source ;
5. Le type de publication.

La collection de tests a été construite dans le cadre d'une étude évaluant l'utilisation de MEDLINE par des médecins en milieu clinique (Hersh et Hickam). Les médecins débutants utilisant MEDLINE ont généré 106 requêtes, un sous-ensemble de ces requêtes a été utilisé dans

TREC-9 [8]. Pour la phase d'évaluation, nous utilisons un sous-ensemble de 63 du jeu de requêtes original développé par Hersh et al [164]. Pour les expériences de recherche d'information (OHSUMED), chaque requête a été répliquée par quatre chercheurs, deux médecins expérimentés en recherche et deux bibliothécaires médicaux. La pertinence des résultats a été évaluée par un groupe de différents médecins, à l'aide d'une échelle en trois points : absolument, éventuellement ou non.

3.4. Organigramme du système

La figure 18 présente l'organigramme du système d'évaluation, dans ce système, nous commençons par indexer le document comme suit :

1. L'exécution de MetaMap sur OHSUMED TREC Corpus, qui permet l'annotation des termes biomédicaux du corpus et étiqueté ces termes par leurs concepts candidats utilisant UML.
2. Le processus de conceptualisation est réalisé avec deux stratégies de conceptualisation telles que mentionnées avant (***Adding Concept, Concept Only***). Pour les termes polysémiques, nous utilisons les stratégies de désambiguïsation :
    a. ***First Concept*** ;
    b. ***All concepts*** ;
    c. Based Context
        i. ***SenseRelate*** ;
        ii. ***NoDistanceSenseRelate***.

Après la création du vecteur descripteur de chaque document composant le modèle d'index, la deuxième étape consiste à calculer le score de pertinente du document par rapport à une requête, qui est réalisée par le système de recherche d'informations Terrier à l'aide de modèles de pondération TF-IDF et BM25.

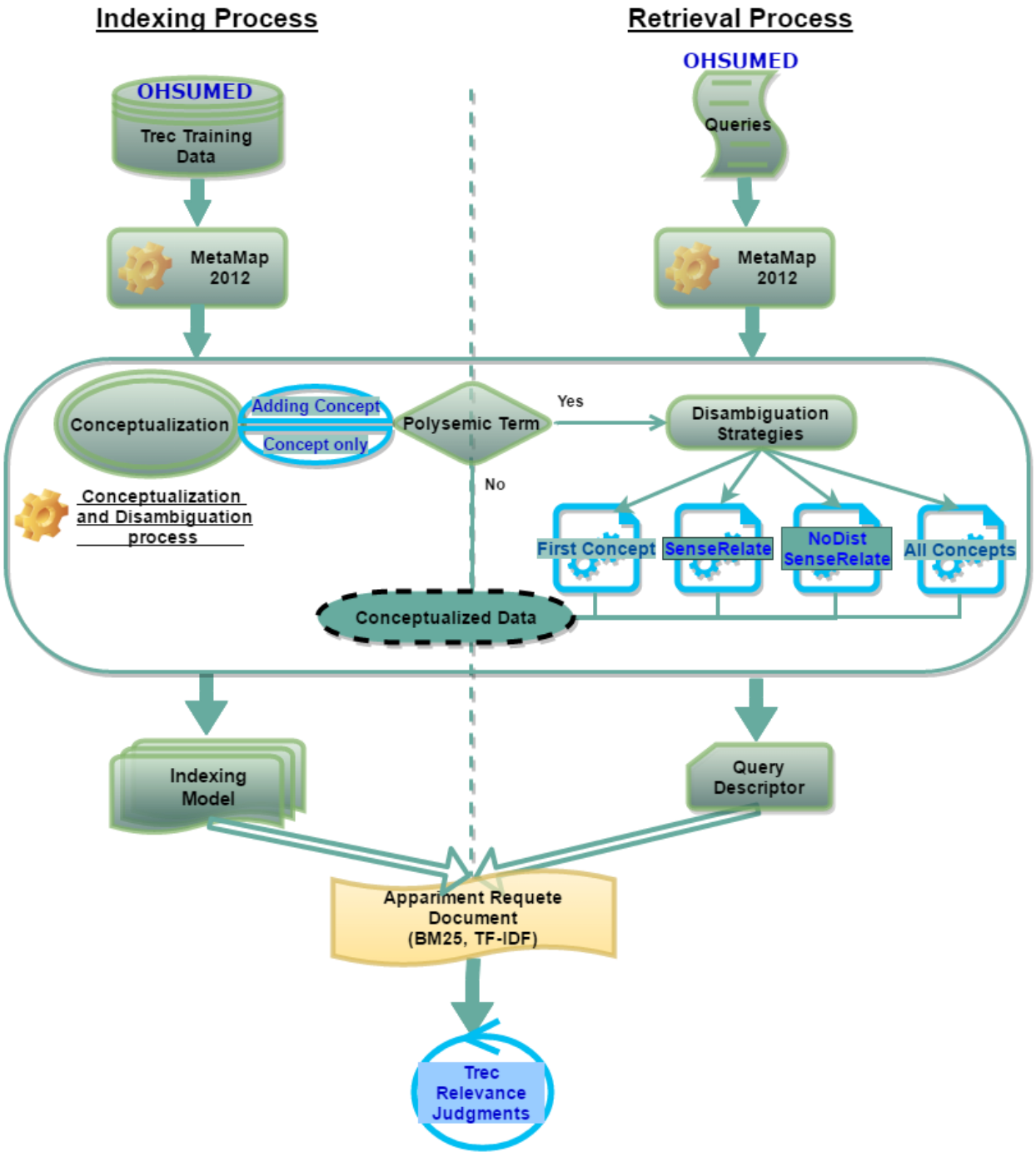


Fig 18 Organigramme du système d'évaluation : Système de recherche d'informations biomédicales

4. Résultats et discussion

Dans cette section, nous présentons les mesures d'évaluation utilisées et les tableaux de résultats obtenus avec la discussion.

4.1. Mesures d'évaluation

A. Précision

Mesure de la capacité d'un système à ne présenter que des éléments pertinents.

$$précision = \frac{nombre\ d'éléments\ pertinents\ récupérés}{nombre\ total\ d'éléments\ récupérés} \quad (6.10)$$

B. Précision en K

Comme de nombreuses requêtes récupèrent des milliers de documents pertinents, peu d'utilisateurs voudront les lire tous. La précision en k (P@k) reste un indicateur utile (par exemple, P@10 ou "Précision en 10" correspond au nombre des résultats pertinents de la première page de résultats de recherche).

C. Moyen de la Précision Moyenne MAP (Mean Average Precision)

La précision moyenne [221] est la moyenne de la valeur de précision obtenue pour l'ensemble des k documents les plus importants, existants après la récupération de chaque document pertinent. Ensuite la moyenne est établie pour cette valeur pour le besoins d'informations.

Moyen de la précision moyenne pour un ensemble de requêtes est la moyenne des scores de précision moyenne pour chaque requête. Nous considérons MAP sur un total de 63 requêtes de test.

D. Classement Réciproque (Reciprocal Rank)

La mesure du classement réciproque [178] [21] favorise les fonctions de score qui accordent un rang élevé aux résultats pertinents. La valeur de la mesure est inversement proportionnelle à la distance parcourue par un utilisateur dans la liste des résultats classés pour trouver le premier résultat pertinent.

E. bpref

La mesure bpref [222] est conçue pour les situations dans lesquelles on sait que les jugements de pertinence sont loin d'être complets, bpref calcule une relation de préférence consistant à savoir si les documents jugés pertinents sont récupérés avant les documents jugés non pertinents. Ainsi, il est basé uniquement sur les rangs relatifs des documents jugés.

4.2. Résultats Expérimentaux

Les tableaux 10 et 11 présentent les résultats obtenus des stratégies de conceptualisation lors de l'application des stratégies de désambiguïsation utilisant respectivement les modèles de pondération TF-IDF et Okapi BM25.

Tableau 10 Résultats des stratégies de conceptualisation et de désambiguïsation en utilisant le modèle TF-IDF

Model TF-IDF							
Conceptualization	**Disambiguation**	**MAP**	**bpref**	**recip_rank**	**P_5**	**P_10**	**P_20**
Term Based		11,13%	53,66%	24,85%	13,97%	12,06%	10,56%
Concept Only	**First Concept**	9,63%	43,93%	24,90%	13,97%	10,79%	8,89%
		-1,50%	-9,73%	*+0,05%*	0,00%	-1,27%	-1,67%
	All Concept	4,09%	38,85%	12,09%	5,71%	4,60%	3,41%
		-7,04%	-14,81%	-12,76%	-8,26%	-7,46%	-7,15%
	NoDistanceSenseRelate	9,73%	42,68%	24,41%	14,60%	10,32%	8,57%
		-1,40%	-10,98%	-0,44%	*+0,63%*	-1,74%	-1,99%
	SenseRelate	10,25%	43,64%	23,89%	15,24%	10,95%	9,05%
		-0,88%	-10,02%	-0,96%	*+1,27%*	-1,11%	-1,51%
Adding Concept	**First Concept**	11,19%	54,34%	24,64%	13,97%	12,70%	11,11%
		+0,06%	*+0,68%*	-0,21%	0,00%	*+0,64%*	*+0,55%*
	All Concept	4,91%	43,57%	13,72%	7,62%	6,19%	4,60%
		-6,22%	-10,09%	-11,13%	-6,35%	-5,87%	-5,96%
	NoDistanceSenseRelate	11,57%	54,41%	26,34%	15,56%	13,17%	10,79%
		+0,44%	*+0,75%*	*+1,49%*	*+1,59%*	*+1,11%*	*+0,23%*
	SenseRelate	12,33%	55,24%	27,61%	15,87%	13,81%	11,35%
		+1,20%	*+1,58%*	*+2,76%*	*+1,90%*	*+1,75%*	*+0,79%*

Tableau 11 Résultats des stratégies de conceptualisation et de désambiguïsation en utilisant modèle BM25

Model BM25							
Conceptualization	**Disambiguation**	**MAP**	**bpref**	**recip_rank**	**P_5**	**P_10**	**P_20**
Term Based		11,16%	53,92%	24,88%	13,97%	12,22%	10,63%
Concept Only	**First Concept**	9,74%	43,93%	25,02%	13,97%	10,79%	8,89%
		-1,42%	-9,99%	***+0,14%***	0,00%	-1,43%	-1,74%
	All Concept	4,20%	39,43%	12,90%	5,71%	4,60%	3,41%
		-6,96%	-14,49%	-11,98%	-8,26%	-7,62%	-7,22%
	NoDistanceSenseRelate	9,86%	43,02%	24,43%	14,60%	10,32%	8,65%
		-1,30%	-10,90%	-0,45%	***+0,63%***	-1,90%	-1,98%
	SenseRelate	10,36%	43,91%	23,90%	15,24%	11,11%	9,21%
		-0,80%	-10,01%	-0,98%	***+1,27%***	-1,11%	-1,42%
Adding Concept	**First Concept**	11,35%	54,07%	25,69%	13,97%	12,86%	11,11%
		+0,19%	***+0,15%***	***+0,81%***	0,00%	***+0,64%***	***+0,48%***
	All Concept	6,02%	51,26%	17,86%	8,89%	7,62%	5,40%
		-5,14%	-2,66%	-7,02%	-5,08%	-4,60%	-5,23%
	NoDistanceSenseRelate	11,64%	54,22%	26,59%	15,56%	13,17%	10,79%
		+0,48%	***+0,30%***	***+1,71%***	***+1,59%***	***+0,95%***	***+0,16%***
	SenseRelate	12,37%	54,98%	27,80%	15,87%	13,97%	11,43%
		+1,21%	***+1,06%***	***+2,92%***	***+1,90%***	***+1,75%***	***+0,80%***

Comme l'illustrent les tableaux 10 et 11, dans tous les cas, nous pouvons observer que la conceptualisation en utilisant la méthode ***Adding Concept*** lors de l'application des stratégies de désambiguïsation ***SenseRelate*** et ***NoDistanceSenseRelate*** améliore le résultat.

L'amélioration des taux de MAP obtenus, illustrés dans les tableaux 10 et 11, est de 1,21% et 1,2% en utilisant respectivement Okapi BM25 et TF-IDF pour ***Adding Concept*** l'hors d'utilisation de l'algorithme ***SenseRelate*** par apport à l'approche basée sur terme. Toutefois l'approche basée sur terme améliore ***SenseRelate*** tout en utilisant la stratégie de ***Concept Only*** avec MAP de 0,8 et 0,88 pour Okapi BM25 et TF-IDF respectivement. En plus, nous pouvons observer que la MAP de la stratégie ***First Concept*** en utilisant ***Adding Concept*** surperforme l'approche basée sur

terme avec une amélioration de 0,19% et 0,06% en utilisant respectivement Okapi BM25 et TFI-DF. Cependant, MAP de la stratégie basée sur terme, présente une amélioration de 1,42 et 1,50 en utilisant respectivement Okapi BM25 et TF-IDF par rapport au First concept, tout en utilisant une stratégie de conceptualisation ***Concept Only***. Cela, démontre l'intérêt de prendre en compte le terme et le concept dans la représentation finale du document à l'aide des stratégies ***Adding Concept***.

Les résultats de la représentation basée sur terme, en général, sont très proches de la performance de ***First Concept***. Nous voyons que ***SenseRelate***, ***NoDistanceSenseRelate***, ***Term Based*** et ***First Concept*** donnent toujours une meilleure précision MAP que la stratégie ***All Concept*** avec ***Adding Concept*** et ***Concept Only***, en raison de l'augmentation de l'espace de représentation.

6. Conclusion

La résolution de l'ambiguïté de sens (WSD) joue un rôle essentiel dans de nombreuses applications de fouille de texte biomédical telles que la recherche d'informations. ***SenseRelate*** est un algorithme bien connu basé sur le contexte WSD. Dans le chapitre 3, nous avons proposé une version simple modifiée de l'algorithme ***SenseRelate*** nommée ***NoDistanceSenseRelate***.

Pour illustrer l'efficacité de ces deux algorithmes, ***SenseRelate*** et ***NoDistanceSenseRelate*** ont été intégrés dans un système d'Indexation et de Recherche d'Informations Biomédicales. Plusieurs expériences ont été menées à l'aide de deux modèles de pondération Okapi BM25 et TF-IDF.

Les résultats obtenus en utilisant le corpus OHSUMED montrent que les méthodes basées sur le contexte (***SenseRelate*** et ***NoDistanceSenseRelate***) sont plus performantes que les autres lors de l'application de la stratégie de conceptualisation ***Adding Concept***, ce qui prouve l'intérêt de l'ajout du sens des concepts dans la Représentation Conceptuelle des documents biomédicaux dans le processus d'indexation.

Chapitre 7 : Navigation et Représentations des Résultats de Recherche de *PubMed*

1. Introduction

Le processus de navigation des résultats de recherche est l'un des problèmes majeurs des moteurs de recherche Web traditionnels. PubMed est un moteur de recherche très populaire dans le domaine des sciences de la vie et de la recherche biomédicale. Il permet un accès gratuit aux bases de données accédant principalement à la base de données MEDLINE contenant des références et des résumés sur les sciences de la vie et les sujets biomédicaux.

Comme tous les autres moteurs de recherche, PubMed a pour objectif de récupérer les citations jugées pertinentes pour une requête utilisateur. Les développeurs de moteurs de recherche modernes ont consacré beaucoup d'efforts à l'optimisation du classement des résultats de recherche, dans l'espoir de placer les plus pertinents en haut de la liste. Néanmoins, aucune solution de classement n'est parfaite, en raison de la complexité inhérente au classement des résultats de recherche [187].

Les aspects de cette complexité proviennent de types de requêtes possibles qui sont très différentes. Des requêtes de sujet restreintes, ou des requêtes de sujet spécifiques, retournant un nombre relativement petit de citations à partir de MEDLINE. La majorité des utilisateurs de PubMed font preuve d'une patience médiocre avec les ensembles de recherche volumineux issus de requêtes de sujets généraux. Ces utilisateurs parcourent généralement la première page ou même les dix premiers résultats dans l'espoir de trouver les bonnes réponses à leurs requêtes. De ce fait, le regroupement des résultats de recherche sur le Web (Web Search Result Clutering - WSRC) est d'une importance cruciale pour le regroupement en ligne de documents similaires afin d'améliorer et de faciliter la navigation sur les pages Web sous une forme plus compacte et thématique.

Pour faire face à ce problème de navigation, de nombreuses solutions de moteurs de recherche ont été proposées ces dernières années, telles que Textpresso [188], XplorMed [189], semedico [192], novo / seek [191] et GoPubMed [190] qui utilisent des vocabulaires contrôlés, tels que Gene Ontology (GO), Medical Subject Headings (MeSH) [193], Systematized Nomenclature of Medicine (SNOMED) et Unified Medical Language System (UMLS) [194], en tant que ressource d'informations pour l'extraction de rubriques à partir des résultats de la recherche.

Dans le cadre de ce travail, nous proposons un nouveau système de navigation. Ce dernier est basé sur le regroupement des résultats de recherches retournées par PubMed. Le système créera des groupes distincts d'extraits Web étiquetés, renvoyés par le moteur de recherche PubMed, afin de répondre aux besoins des utilisateurs de la scientifique biomédicale. Dans ce nouveau système de regroupement des résultats de recherche Web des documents biomédicaux, nous utilisons l'analyse formelle de concept (Formal Concept Analysis - FCA) [195]. La FCA a été utilisée avec

succès comme nouvelle méthode de classification des résultats de recherche Web basée sur la classification conceptuelle. Elle a été intégrée dans de nombreux systèmes pour résoudre le problème de la navigation sur le Web, en particulier pour les langues [196] [197] [198].

Le reste de ce chapitre est structuré comme suit : le contexte général et la tâche de regroupement des recherches sur le Web sont présentés dans la section 2. La section 3 décrit les méthodes et l'architecture utilisées pour notre système d'évaluation. Les résultats et discussion de notre système d'évaluation se trouvent dans la section 4 et à la fin nous concluons ce chapitre.

2. Contexte de l'étude

2.1. Regroupement des résultats de recherche WEB

Le regroupement des résultats de recherche **(Web Search Results Clustering – WSRC)** vise à organiser les extraits partageant un sujet commun dans un même cluster et à former des étiquettes correspondantes pour la description du cluster. Une telle technique est développée pour répondre au problème de surcharge d'information : les utilisateurs sont souvent submergés par une longue liste de documents retournés. Récemment, il est devenu l'un des domaines centraux de la recherche pour résoudre le problème de la navigation sur le Web, et de nombreuses approches ont été proposées qui peuvent être classées en deux catégories : Data-Centric et Description-Centric. Pour plus de détails, dans ce cadre le travail de Carpineto et *al*.,[199] a présenté une enquête.

A. L'approche : Data-Centric

L'approche data-centric - centrée sur les données - regroupe un ensemble de systèmes WSRC basés sur des algorithmes de classification classiques tels que Hierarchical [200], K-means [201] et Spectral [202], qui sont appliqués aux résultats de recherche de groupe et souvent légèrement adaptés pour produire une description de cluster significative. Cette catégorie contient de nombreux exemples de systèmes tels que Lassi [203], CIIRarchies [204], Armil [205] et Scatter / Gather [206]. De manière générale, le problème le plus critique de toutes les approches de cette catégorie est la qualité des étiquettes du cluster. En fait, lorsque la qualité des étiquettes d'un cluster est prioritaire dans les résultats de la recherche de cluster, la deuxième catégorie devient plus importante pour produire des groupes avec des étiquettes compréhensibles. Ces étiquettes ne sont pas sélectionnées de manière aléatoire, mais elles doivent être liées au sujet étudié.

B. Approche Description-Centric

La première méthode de cette catégorie a été proposée par Zamir et *al.* Elle a été nommée Grouper [207]. Il s'agit d'une technique de classification en ligne basée sur la structure de données d'arborescence de suffixe, dans laquelle les résultats de la recherche sont regroupés et les grappes étiquetées à l'aide des expressions courantes trouvées par la structure de données d'arborescence de suffixe. Une autre solution de cette catégorie, la FCA, qui est une théorie mathématique introduite par Ru-dolf Wille en 1984, a été intégrée dans de nombreux systèmes du WSRC, tels que JBreanDead [197], Credo [196] et CHC [198]. Bien que la FCA a été utilisée avec succès comme technique de classification conceptuelle pour résoudre le problème de WSRC avec une description intentionnelle de chaque cluster afin de rendre les regroupements facilement interprétables, son principal inconvénient est que le treillis de concepts généré peut être ingérable lorsqu'il est appliqué à de grandes collections de documents et de riches ensembles de termes d'indexation [208] [209] [210] [211] [212]. Dans cet article, nous étudions comment intégrer la FCA dans un nouveau système pour Biomédical WSRC.

2.2. Systèmes basés sur PubMed

Il existe de nombreux systèmes qui présentent les résultats en cluster dans différentes étiquettes ou catégories à l'aide de documents renvoyés par PubMed. Anne O'Tate [213] post-traite les résultats de recherches PubMed et les regroupes dans l'une des catégories prédéfinies : mots importants, thèmes MeSH, affiliations, noms d'auteurs, revues et année de publication.

McSyBi [214] présente les résultats en cluster de deux manières distinctes : hiérarchique et non hiérarchique. Alors que le premier fournit une vue d'ensemble des résultats de la recherche, le dernier montre les relations entre les résultats de la recherche. En outre, il permet aux utilisateurs de regrouper leurs résultats en leur imposant soit un terme MeSH, soit le type ULMS Sémantique de son domaine de recherche. Les groupes mis à jour sont automatiquement étiquetés par les termes MeSH pertinents et par les termes de signature extraits du titre et des résumés. GOPubMed [190] a été initialement conçu pour exploiter la hiérarchie de Gene Ontology (GO) afin d'organiser les résultats de la recherche, permettant ainsi aux utilisateurs de naviguer rapidement dans les résultats par catégories. XplorMed [215] organise non seulement les résultats par classes MeSH, mais permet également aux utilisateurs d'explorer le sujet et les mots qui les intéressent.

Dans cet article, nous étudions comment intégrer FCA dans un nouveau système pour Biomédicale WSRC utilisant le moteur de recherche PubMed.

2.3. Analyse de Concepts Formels (Formal Concept Analysis - FCA)

L'idée de base du modèle FCA consiste à explorer d'abord le contexte formel entre les extraits d'objets classés obtenus par les moteurs de recherche, puis de construire le Treillis de concepts en tant que représentation des nouveaux extraits. Dans cette section, nous présentons la théorie de l'analyse formelle de concept en donnant quelques définitions importantes et quelques exemples illustratifs.

A. Contexte Formel (G, M, I)

Le contexte formel (G, M, I) est constitué d'un ensemble d'objets G, d'un ensemble d'attributs M et I est défini par une relation binaire entre les objets G et les attributs M dans un ensemble de données qui relie les objets aux valeurs des attributs. Le tableau 12 montre un exemple de contexte formel.

B. Concept Formel du Contexte Formel (G, M, I)

Le concept formel du contexte formel (G, M, I) est un ensemble d'objets partageant des caractéristiques similaires. En utilisant la définition mathématique donnée par Rudolf Wille, le concept formel est défini comme un couple (A, B) avec A # G, B # M, $A = B^I$ et $B = A^I$. A et B sont respectivement appelés l'étendue et l'intention du concept formel (A, B) [195].

Où:

$$A^I = \{m \in M \mid gIm \ \forall g \ \in A\} \tag{7.1}$$

$$B^I = \{g \in G \mid gIm \ \forall m \in B\} \tag{7.2}$$

Avec :

- A^I est l'opérateur de dérivation de A ;
- B^I est l'opérateur de dérivation de B.

C. Treillis de Concepts (G, M, I)

Treillis de concepts (***Concept lattice***) (G, M, I) est une hiérarchie ordonnée de tous les concepts formels du contexte formel (G, M, I). De nombreux algorithmes ont été proposés pour construire le Concept Lattice à partir du contexte formel. Ils peuvent être classés en deux catégories : (a): Les algorithmes sont développés pour améliorer les performances en générant un ensemble de concepts tels que Ganter [216]; (b): Des algorithmes sont développés pour améliorer les performances dans la construction de tout le réseau, tels que ceux de Godin's et *al.* [217], Bordat [218] et Nourine et Raynaud [219]. La Figure 19 montre le concept lattice correspondant au contexte formel présenté dans le tableau 12.

Tableau 12 Exemple de contexte formel

	Protein	***Arthropod***	***Human***	***Tick-***	***West nile***	***Yellow Fever***	***secretin test***
G1	1	1	0	0	0	0	0
G2	1	1	0	1	0	0	0
G3	0	1	0	0	1	0	0
G4	1	0	0	0	0	1	0
G5	1	0	1	0	0	0	0
G6	1	1	0	0	0	0	0
G7	0	0	1	1	0	0	0
G8	1	0	0	0	0	1	1
G9	1	1	0	0	1	0	0

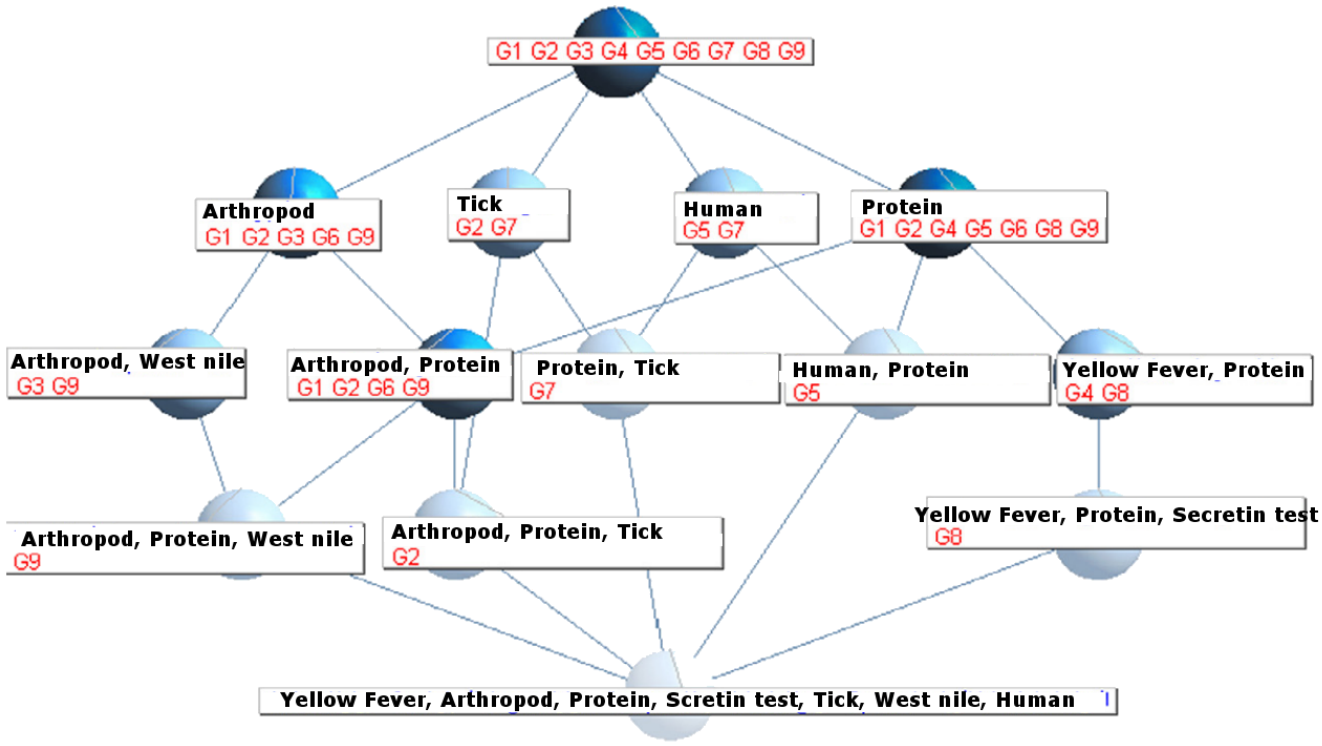


Fig 19 Treillis de concepts correspondant au contexte formel du Tableau 12

3. Méthode d'évaluation

Dans cette section, nous présentons en détail notre système proposé pour le regroupement des résultats de la recherche biomédicale sur le Web, basé sur l'analyse formelle de concept.

3.1. Organigramme proposé

L'organigramme de notre nouveau système proposé est présenté dans la Figure 20.

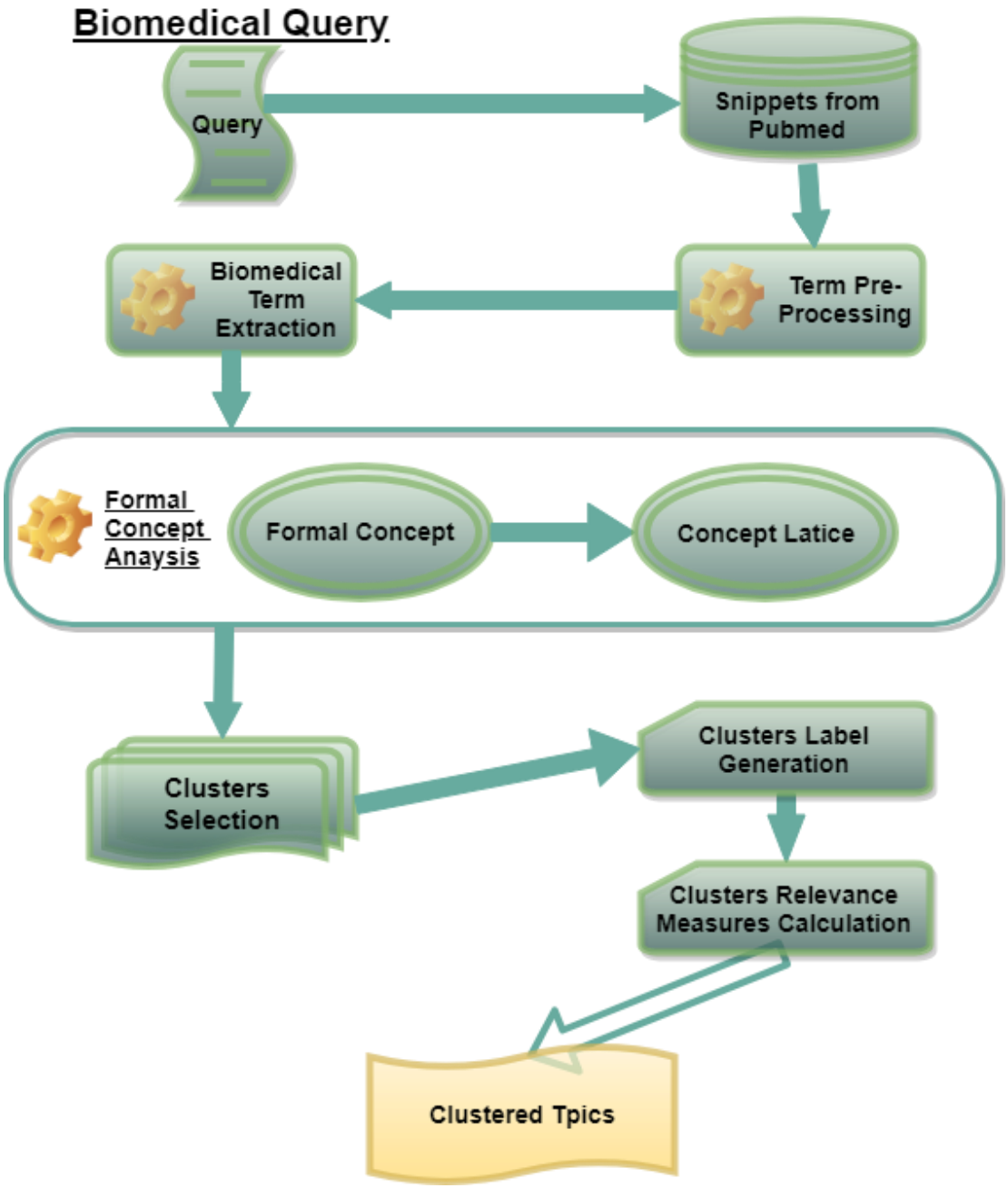


Fig 20 Organigramme des résultats de système de recherche d'information biomédicale

L'organigramme de notre système est résumé par les étapes suivantes :

1. Récupération des extraits depuis PubMed en envoyant une requête biomédicale à E-utilities [7] ;
2. Exécution de module de prétraitement du texte et extraction de termes biomédicaux pour extraire les termes biomédicaux tout en créant des vecteurs de termes pour chaque extrait ;
3. Construction du Concept lattice ;
4. Sélection des clusters ;
5. Génération des étiquettes des clusters.

3.2. Construction de contexte formel

Les utilisateurs de ***PubMed*** spécifient leur requête avec des termes biomédicaux à l'aide de l'interface Web de nos systèmes. La requête est envoyée aux moteurs de recherche Web ***PubMed*** à l'aide du service électronique proposé par ***PubMed*** E-utility. La liste des résultats renvoyés est sous forme d'extraits, afin de rendre notre implémentation simple et facile, à chaque extrait, nous associons les quatre balises suivantes (ID, URL, Corps et Titre).

Où:

- ID: identifiant du fragment de document.
- URL: lien pour accéder au contenu du document.
- Corps: contenu de l'extrait.
- Titre: titre de la page.

Chaque extrait est nettoyé en supprimant les Mots Vides (***Stop Words***) anglais et les caractères spéciaux tels que, (/, 'n #, n $, etc.,…). Après cela, nous exécutons le module Term Extractor du domaine biomédical, qui tente d'extraire les termes biomédicaux du texte en utilisant UMLS comme dictionnaire. Enfin, les termes obtenus représentent l'ensemble des attributs et les extraits avec ID représentent l'ensemble des objets dans le contexte formel.

3.3. Construction du Concept Lattice et Sélection de Clusters

Le contexte formel obtenu est utilisé pour construire le Concept Lattice. Dans notre cas, nous utilisons l'API Java gratuite nommée ToscanaJ, qui intègre l'algorithme de Ganter [216] pour générer l'ensemble des concepts formels et le Concept Lattice correspondant.

Ce dernier représente un ensemble de concepts organisés en structure hiérarchique, chaque concept regroupant un ensemble de documents (objets représentés par les ID de fragments dans des lignes de contexte formel) qui représentent l'étendue partageant un ensemble de termes (attributs représentés par des termes dans des colonnes de contexte formel) qui représentent l'intention.

Notez que le concept représente un cluster lors de l'utilisation de FCA pour un processus de cluster [196] [197] [198]. Dans ce travail, nous proposons d'utiliser les concepts obtenus qui apparaissent aux premier et deuxième niveaux de la hiérarchie de Concept Lattice en tant que clusters sélectionnés. En fait, nous sélectionnons uniquement les premiers et deuxièmes niveaux pour obtenir plus de clusters séparés avec des étiquettes plus descriptives. De plus, afin de faciliter la navigation de l'utilisateur, les clusters obtenus doivent être mappés du Concept Lattice vers une interface utilisateur graphique et présentés en fonction de leur pertinence pour la requête de l'utilisateur. À cette fin, nous avons développé une interface utilisateur graphique simple et efficace dans laquelle les clusters obtenus sont classés en tenant compte de leur pertinence pour la requête de l'utilisateur.

En règle générale, le problème du classement de la pertinence de la grappe est d'estimer la pertinence d'un concept correspondant à la requête d'un utilisateur. Pour surmonter ce problème, Zhang et *al.* [198] ont proposé une nouvelle méthode pour construire la hiérarchie à deux niveaux réduits à partir du Concept Lattice. Cette méthode repose sur deux mesures mathématiques: la première est la mesure de l'importance du concept utilisée pour indiquer son importance. Cette mesure concerne à la fois le nombre de documents dans l'étendue et le nombre de concepts descendants de ce concept. La seconde mesure est la similarité de concept, basée sur le coefficient de similarité de Jaccard et sera utilisée dans le processus de fusion pour construire une hiérarchie à deux niveaux pour la navigation de l'utilisateur.

Le système créé par Zhang et Feng repose sur l'utilisation de tous les termes déduit à partir des extraits. Par conséquent, un processus de réduction est nécessaire pour réduire un certain nombre de grappes générées non significatives en filtrant ou en regroupant des concepts similaires. Par ailleurs, nous utilisons des termes nominatifs au lieu d'utiliser tous les termes sans qu'il en soit ainsi, ce qui rend la réduction inutile. En fait, le nombre de termes nominaux dans chaque extrait est très faible. En outre, ils sont liés au sujet du contenu du document correspondant. Cependant, l'estimation de la pertinence d'un concept correspondant à la requête d'un utilisateur est nécessaire pour faciliter le processus de navigation de celui-ci. Comme nous l'avons mentionné ci-dessus, un concept est caractérisé par deux composants : l'étendue et l'intention.

Par conséquent, dans ce travail, nous proposons notre nouveau concept de mesure de la pertinence, qui prend en compte les deux composants suivants :

- le nombre de documents dans l'étendue ;
- Le poids de chaque mot dans l'intention.

3.4. Génération de label de cluster

La génération d'étiquettes de grappes est une étape cruciale, car les étiquettes sans signification ou trompeuses peuvent amener les utilisateurs à vérifier les mauvaises grappes. De plus, les étiquettes devraient être plus compréhensibles pour les utilisateurs et décrire avec précision le contenu des documents. À cette fin, nous proposons de trouver le terme d'origine de chaque élément dans l'intention du concept correspondant. Le libellé du cluster est un ensemble de termes d'origine séparés par une virgule. Ensuite, l'utilisateur peut simplement cliquer sur l'étiquette du cluster produit qui décrit aux mieux ses besoins spécifiques en informations dans la hiérarchie des thèmes.

4. Résultats et discussion

Dans cette section, nous présentons une étude comparative entre notre système proposé et deux autres comme base : STC et Lingo. STC est un WSRC classique basé sur une structure de données d'arbre de suffixe et Lingo est un algorithme bien connu du WSRC dans le domaine universitaire. Les deux systèmes sont intégrés à la plate-forme *Carrot2*, qui est un moteur de regroupement des résultats de recherche open source. Cette étude comparative s'intéresse à la qualité des résultats de regroupement et du label produit par différents systèmes.

4.1. Corpus

Le corpus TREC Genomics 2007 [220] est utilisé afin d'évaluer notre méthode. Les principales caractéristiques de ce corpus sont les liaisons et les annotations. La liaison entre les ressources permet à l'utilisateur d'explorer différents types de connaissances parmi ces ressources. De plus, il fournit une annotation de thème pour chacun des documents

4.2. Regroupement et la qualité des résultats

En règle générale, la qualité des résultats de regroupement de tout système de regroupement peut être mesurée par son degré de capacité à reclasser correctement un ensemble d'extraits préclassés dans exactement les mêmes catégories sans connaître l'affectation de catégorie d'origine. La qualité des résultats de regroupement peut être mesurée par deux métriques ; l'Information mutuelle normalisée (NMI) et Entropie complémentaire normalisée (NCE). Ces métriques sont utilisées par Geraci et *al.* comparer l'efficacité de différents algorithmes WSRC [205]. Pour une série donnée S de N extraits pré-classifiés sous C = {c1, c2, ..., cn} de catégories et un ensemble C '= {c'1, c'2, ..., c'm} comme les résultats du regroupement NMI et NCE sont définis comme suit :

$$NMI(C, C') = \frac{2}{\log|C||C'|} \sum_{c \in C} \sum_{c' \in C'} P(c, c') \log \frac{P(c, c')}{P(c)P(c')} \tag{7.3}$$

Où

$$P(c) = \frac{|c|}{N},\ P(c') = \frac{|c'|}{N},\ P(c, c') = \frac{|c| \cap |c'|}{N} \tag{7.4}$$

$$NEC(C, C') = \sum_{i=1}^{m} \frac{|c'_i|}{N'} NCE(C, c'_i) \tag{7.5}$$

Où

$$NEC(C, c'_i) = 1 - \frac{2}{\log|C|} \sum_{j=1}^{n} - \frac{P(c_j, c'_i)}{P(c_j)} \log \frac{P(c_j, c'_i)}{P(c_j)} \tag{7.6}$$

$$N' = \sum_{i=1}^{m} |c'_i| \tag{7.7}$$

NMI est conçu pour un regroupement sans chevauchement. Par conséquent, ses valeurs plus élevées indiquent une meilleure qualité de regroupement. NCE se situe dans l'intervalle [0, 1] et est conçu pour prendre en compte le chevauchement, une valeur plus grande de NCE signifie une meilleure classification.

Zhang et Feng [198] ont montré que ces métriques souffrent de biais pour les raisons suivantes:

- Si les catégories d'origine sont fixes, plus le nombre de clusters générés par un algorithme WSRC donné est élevé, plus les valeurs de NMI et de NCE obtenues sont élevées.
- Si les groupes à comparer sont fixes, plus le nombre de groupes dans les catégories d'origine est élevé, plus les valeurs de NMI obtenues sont élevées.
- Lors de la comparaison des clusters résultants générés par deux algorithmes WSRC différents avec NCE et NMI, les performances peuvent être inversées si des catégories d'origine différentes sont utilisées.

Pour surmonter les préjugés susmentionnés des deux métriques, Zhang et Feng ont proposé deux métriques améliorées: A-NMI@K et A-NEC@K, où A indique la moyenne de chaque résultat des catégories utilisées et K indique le nombre de clusters sélectionnés pour l'expérience. L'utilisation des métriques améliorées prend en compte la variation des résultats expérimentaux lors du changement de catégorie et peut donner une idée globale des performances des systèmes utilisés pour l'expérience. Dans cette étude comparative, nous utilisons uniquement K = 10, K = 20 et K =

all, car nous considérons que 10 est le nombre minimal de grappes visionnées par un utilisateur et 20 est le maximum. De plus, les résultats restent les mêmes pour K=5 et K=15.

La figure 21 présente A-NMI@10, A-NMI@20 et A-NMI@ALL pour des mesures de qualité de regroupement sans chevauchement des trois systèmes : Notre système, STC et Lingo. Il est clair que notre système surpasse les deux autres et offre une amélioration deux fois supérieure par rapport aux deux autres systèmes.

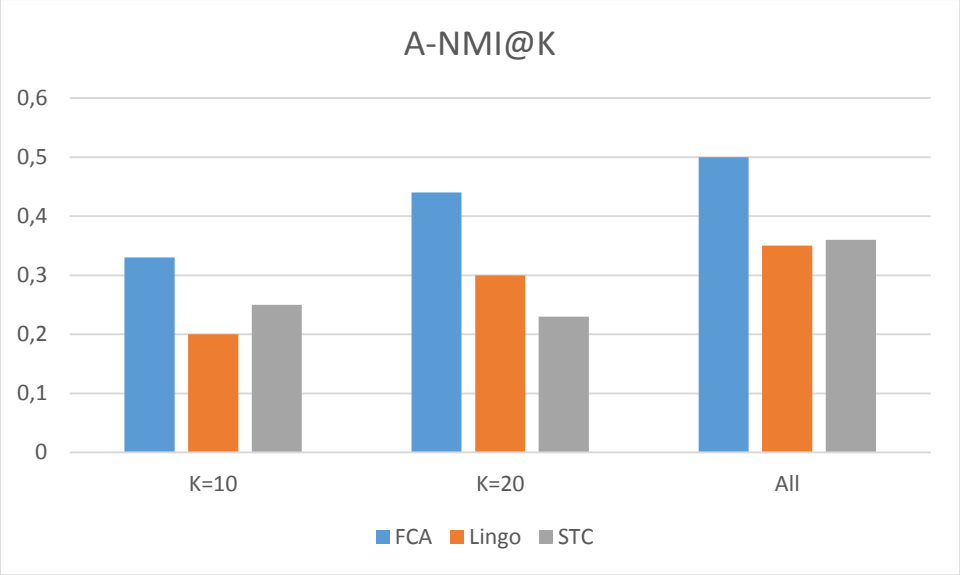


Fig 21 A-NMI@K pour les mesures de qualité de regroupement ne se chevauchant pas

5. Conclusion

La navigation dans les résultats de recherche est l'un des problèmes majeurs des moteurs de recherche traditionnels (Google, Yahoo et Bing). En ce qui concerne les sciences de la vie et la recherche biomédicale, les moteurs de recherche les plus utilisés est PubMed, qui présentent le même problème et dont le style de navigation semble donc ne pas être convivial. Dans ce chapitre, nous avons présenté l'utilisation de l'analyse formelle de concept dans un nouveau système de WSRC du domaine biomédical. Le système proposé regroupe automatiquement les résultats de la recherche Web en groupes de haute qualité dotés d'une structure hiérarchique, et fournit des étiquettes descriptives de groupe. Une série d'expériences a été menée à l'aide du service Web E-utilities. Les résultats obtenus ont été très encourageants et illustrent l'efficacité du système proposé.

Conclusion Générale et Perspectives

Le volume des publications dans le domaine biomédical s'accroit à un rythme considérable : plus de 23 millions de citations publiées recensées dans la base MEDLINE et disponibles via PubMed. Notons aussi qu'il y a une augmentation quotidienne de 1800 références. Dans cette masse de données textuelles, la recherche manuelle d'information biomédicale est très difficile et non efficace. De ce fait, l'exploitation de ces ressources devient impossible.

Les travaux présentés dans cette thèse se situent dans le contexte précis de l'accès à l'information médicale. Plus précisément, nos contributions portent sur les quatre volets ou problèmes suivants :

- **Problème de reconnaissance d'entités nommées biomédicales [Bio NER].**
- **Intégration de la Sémantique dans la représentation du texte biomédical et la résolution de l'ambigüité lexicale des termes biomédicaux.**
- **Problème d'indexation des documents biomédicaux.**
- **Problème de navigations et de représentations des résultats de recherche des documents biomédicaux.**

Pour répondre et remédier à ces questions, nous avons proposé des contributions à travers des études liées à la représentation et à l'indexation des documents biomédicaux basées sur l'utilisation des ressources termino-ontologiques. Puis, nous avons mené plusieurs expérimentations pour montrer l'efficacité de nos études et nos propositions.

Dans le cadre de cette thèse et pour remédier au problème de la reconnaissance des entités nommées biomédicale, nous avons élaboré une étude comparative d'une part théorique, et d'autre part expérimentale entre sept méthodes basées sur l'approche d'apprentissage automatique, il s'agit des méthodes suivantes : (CRF, MEMM, HMM, Support Vector Machine SVM, Naïve Bayes NB, Maximum Entropy ME, et Decision Tree DT). Cette étude très riche, nous permet par la suite de justifier le choix objectif de la méthode Conditional Random Fields CRF qui a produit le meilleur résultat en termes de rappel, précision, et F1-measure.

Pour remédier au problème de l'intégration de la sémantique dans la représentation du texte biomédical et la résolution de l'ambigüité lexicale des termes biomédicaux, nous avons bien mené une étude théorique qui porte sur la représentation du texte biomédical, dont nous avons discuté les différentes stratégies de conceptualisation et de résolution de l'ambiguïté et présenté un cadre de travail générique pour la conceptualisation de texte biomédical. Ensuite, nous avons réalisé une étude sur l'impact de la distance des termes du contexte sur l'algorithme **SenseRelate** [112]. Pour y parvenir, nous avons apporté quelques modifications à SenseRelate de manière à ce

que le score final du terme en cours d'analyse ne soit pas basé sur la distance entre lui-même et les termes de son contexte, ce qui est peut être considéré comme une nouvelle méthode que nous avons nommés **NoDistanceSenseRelate**. De plus, nous avons effectué une étude comparative entre les deux algorithmes (**SenseRelate** et **NoDistanceSenseRelate**), en comparant leurs performances dans un corpus standard biomédical (NLM's MSH-WSD) qui contient 203 termes ambigües.

Par la suite, nous avons proposé une nouvelle mesure de similarité pour le texte court basée sur la théorie des ensembles approximatifs. Cette mesure a été intégrée dans un algorithme de désambiguïsation des termes biomédicaux en utilisant ***PubMed*** comme étant une grande ressource d'information biomédicale. Aussi, nous avons évalué la méthode proposé en comparant ses performances avec d'autres méthodes dans le corpus (NLM's MSH-WSD).

Pour remédier au problème de la représentation des documents du domaine biomédical nous avons proposé d'utiliser une méthode basée sur l'approche conceptuelle qui consiste à représenter un document par un vecteur de concepts (***Bag of Concepts - BoC***) au lieu et à la place d'un vecteur de mots (***Bag of Words - BoW).*** Cette dernière méthode, consiste à intégrer un block pour la désambiguïsation lexicale des termes ambigus.

Dans l'objectif d'évaluer la solution proposée, nous avons commencé d'abord par la réalisation d'un système de catégorisation des documents biomédicaux, tout en intégrant trois stratégies de désambiguïsation lexicale des termes biomédicaux ambigus : ***First Concept, All Concept, Context-Based*** dans le processus de la représentation du texte biomédical, De plus, nous avons bien mené une étude comparative entre ces dernières stratégies à travers une série d'expérimentations en se basant sur l'ontologie du domaine biomédical UMLS en utilisant les méthodes d'apprentissage automatique (ME, NB et SVM), dans un corpus adopté par les spécialistes du domaine : Ohsumed [164].

Pour faire face au problème d'indexation, nous avons effectué une étude théorique concernant la résolution de l'ambiguïté lexicale dans les systèmes de recherche d'information biomédicale, ensuite nous avons intégré notre solution proposée de la représentation conceptuelle basée sur la désambiguïsation lexicale des termes biomédicaux en utilisant les deux méthodes (***SenseRelate***, ***NoDistanceSenseRelate)*** dans un processus d'indexation sémantique des documents biomédicaux. De plus, nous avons bien mené une étude évaluative du système de recherche d'information basé sur l'indexation sémantique proposé à travers une série d'expérimentations sur 63 requêtes (chacune est fournie avec un ensemble de documents jugés pertinents par un groupe de médecins) de la collection OHSUMED, en utilisant deux modèles de recherche d'information ***BM25*** et ***TF_IDF***.

Pour le problème de navigation, nous avons réalisé une étude sur le problème de représentions des documents biomédicaux dans le moteur de recherche ***PubMed*** pour faciliter l'accès à l'information biomédical, ensuite nous avons proposés un système basé sur l'Analyse formelle de concepts - FCA pour le regroupement des résultats de recherche d'information biomédicale en fonction de leur structure hiérarchique.

Perspectives

Dans la continuité directe de notre travail de thèse, nous envisageons plusieurs perspectives à différents niveaux, dont nous citons ci-dessous celles qui nous paraissent les plus prometteuses à savoir :

Nous envisageons comme première perspective de mener une étude plus détaillée sur les deux algorithmes de désambiguïsation SenseRelate et web-kernel en comparent leurs performances dans d'autres corpus plus récents du domaine biomédical et aussi évaluer leurs performances en termes de temps d'exécutions afin de sortir par des conclusions plus précises.

Comme deuxième perspectives, nous envisageons d'évaluer et comparer les deux algorithmes de WSD cités, dans un système de recherche d'information basé sur une indexation sémantique contenant un bloc de WSD qui comprend les deux algorithmes. Et Ainsi présenter le système de recherche d'information proposée dans une application WEB qui se base sur des collections médicales au choix et aussi qui donne le choix d'utilisation ou non d'un bloc de WSD dans le modèle d'index.

Concernant le problème de navigation dans ***PubMed,*** nous pensons qu'il serait possible d'améliorer les performances du système proposé en réduisant le contexte formel textuel, ainsi nous envisageons d'améliorer la génération des labels du premier niveau de navigation. Et finalement publiés l'application web, afin de concevoir une étude subjective sur l'expérience des utilisateurs.

Bibliographie

[1]. Zweigenbaum P, Demner-Fushman D, Yu H, et Cohen KB, 2007. Frontiers of biomedical text mining: Current progress. Briefings in Bioinformatics.; 8(5):358-375

[2]. Luscombe NM, Greenbaum D, et Gerstein M., 2001. What is bioinformatics? A proposed definition and overview of the field. Methods of Information in Medicine. 40(4):346-358

[3]. Krallinger M, Erhardt RA, et Valencia A, 2005, Text-mining approaches in molecular biology and biomedicine. Drug Discovery Today. 10(6):439-445 Review

[4]. D. R. Swanson, 1986, Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30(1):7–18.

[5]. U.S NLM, "Welcome to Medical Subject Headings!" https://www.nlm.nih.gov/mesh/meshhome.html

[6]. U.S NLM, "MEDLINE®: Description of the Database" https://www.nlm.nih.gov/bsd/medline.html

[7]. Eric Sayers, "Entrez Programming Utilities Help", NCBI Help Manual, Bethesda (MD): National Center for Biotechnology Information (US) http://www.ncbi.nlm.nih.gov/books/NBK25501/.

[8]. TREC-9 filtering track collections.http://trec.nist.gov/data/t9_filtering.html.

[9]. TREC genomics track data. http://ir.ohsu.edu/genomics/data.html.

[10]. J.-D. Kim, T. Ohta, Y. Tateisi, et J. Tsujii, 2003, GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics, 19(Suppl 1):i180–i182.

[11]. V. Vincze, G. Szarvas, R. Farkas, G. Mora, et J. Csirik, 2008, The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics, 9(Suppl 11):S9.

[12]. L. Hirschman, A. Yeh, C. Blaschke, et A. Valencia, 2005, Overview of BioCreAtIvE: Critical assessment of information extraction for biology. BMC Bioinformatics, 6(Suppl 1):S1.

[13]. M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, et A. Valencia, 2008, Evaluation of textmining systems for biology: Overview of the second BioCreAtIvE community challenge. Genome Biology, 9(Suppl 2):S1.

[14]. M. Liberman, M. Mandel, et GlaxoSmithKline Pharmaceuticals R&D, 2008. PennBioIE CYP 1.0, Linguistic Data Consortium, Nov 2008.

[15]. M. Liberman, M. Mandel, et P. White, 2008. PennBioIE Oncology 1.0.

[16]. CALBC challenge. http://www.calbc.eu/

[17]. HighWire press. http://highwire.org/

[18]. H. Muller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, C. E. Charles E. Kahn, Jr., et W. Hersh, 2010. Overview of the clef 2010 medical image retrieval track. In Working Notes of CLEF 2010.

[19]. J. Kalpathy-Cramer, H. M¨ uler, S. Bedrick, I. Eggel, A. de Herrera, et T. Tsikrika, 2011. The CLEF 2011 medical image retrieval and classification tasks. InCLEF 2011 Working Notes.

[20]. CRAFT: The colorado richly annotated full text corpus. http://bionlp-corpora.sourceforge.net/CRAFT/index.shtml.

[21]. PubMed central open access subset. http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.

[22]. M. Saeed, M. Villarroel, A. Reisner, G. Clifford, L. Lehman, G. Moody, T. Heldt, T. Kyaw, B. Moody, et R. Mark, 2011. Multiparameter intelligent monitoring in intensive care II (MIMICII): A public-access intensive care unit database. Crit Care Med, 39(5):952–960.

[23]. University of Pittsburgh NLP repository. http://www.dbmi.pitt.edu/nlpfront.

[24]. O. Uzuner, 2009. Recognizing obesity and comorbidities in sparse data. Journal of the American Medical Informatics Association, 16(5):561–570.

[25]. O. Uzuner, I. Goldstein, Y. Luo, et I. Kohane, 2008. Identifyingn patient smoking status from medical discharge records. Journal of the American Medical Informatics Association, 15(1):14–24.

[26]. O. Uzuner, I. Solti, et E. Cadag, 2010. Extracting medication information from clinical text.Journal of the American Medical Informatics Association, 17(5):514–518.

[27]. O. Uzuner, B. R. South, S. Shen, et S. L. DuVall, 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text.Journal of the American Medical Informatics Association, 18(5):552–556.

[28]. G. Leech, 2005. Adding linguistic annotation. In M. Wynne, editor, Developing Linguistic Corpora: A Guide to Good Practice, pages 17–29. Oxbow Books.

[29]. J. W. Wilbur, A. Rzhetsky, et H. Shatkay, 2006. New directions in biomedical text annotation: Definitions, guidelines and corpus construction. BMC Bioinformatics, 7:356.

[30]. H. Shatkay, J. W. Wilbur, et A. Rzhetsky, 2005. Annotation guidelines, http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/AnnotationGuidelines.pdf.

[31]. R. Artstein et M. Poesio, 2008. Inter-coder agreement for computational linguistics.Computational Linguistics, 34(4):555–596.

[32]. P. V. Ogren, 2006. Knowtator: A prot´ eg´ e plug-in for annotated corpus construction. InProceedings of the2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 273–275.

[33]. eHOST: The extensible human oracle suite of tools. http://code.google.com/p/ehost/.

[34]. D. Maynard, 2007. D1.2.2.1.3 benchmarking of annotation tools, 2007. http://knowledgeweb.semanticweb.org/semanticportal/deliverables/D1.2.2.1.3.pdf.

[35]. S. Dipper, M. G¨ otze et M. Stede, 2004. Simple annotation tools for complex annotation tasks: An evaluation. InProceedings of the LREC Workshop on XML-Based Richly Annotated Corpora, pages 54–62.

[36]. D. A. Lindberg, B. L. Humphreys, et A. T. McCray, 1993. The unified medical language system.Methods of Information in Medicine, 32(4):281–291.

[37]. Databases, resources & APIs. http://wwwcf2.nlm.nih.gov/nlm_eresources/eresources/search_database.cfm.

[38]. Open biological and biomedical ontologies. http://www.obofoundry.org/.

[39]. National center for biomedical ontology. http://www.bioontology.org/.

[40]. NCBO BioPortal. http://bioportal.bioontology.org/

[41]. National centre for text mining. http://www.nactem.ac.uk/

[42]. European bioinformatics institute. http://www.ebi.ac.uk/

[43]. Pharmacogenomics knowledge base. http://www.pharmgkb.org/

[44]. Neuroscience information framework. http://neuinfo.org/

[45]. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cheryy, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, et G. Sherlock, 2000. Gene ontology: Tool for the unification of biology. Nature Genetics, 25(1): 25–29.

[46]. M. Q. Stearns, C. Price, K. A. Spackman, et A. Y. Wang, 2001. SNOWMED clinical terms: Overview of the development process and project status. InProceedings of the AMIA Symposium, pages 662–666.

[47]. Orange book: Approved drug products with therapeutic equivalence evaluations. http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm.

[48]. Orange book: Approved drug products with therapeutic equivalence evaluations. http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm

[49]. A. R. Aronson et F.-M. Lang, 2010. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3): 229–236.

[50]. B. Settles, 2005. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics, 21(4): 3191–3192.

[51]. R. Leaman et G. Gonzalez, 2008. BANNER: An executable survey of advances in biomedical named entity recognition. InPacific Symposium on Biocomputing, pages 652–663.

[52]. R. Kabiljo, A. B. Clegg, et A. J. Shepherd, 2008. A realistic assessment of methods for extracting gene/protein interactions from free text. BMC Bioinformatics, 10:233.

[53]. H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Daml-janovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, et W. Peters, 2011. Text Processing with GATE (Version 6), Gateway Press CA.

[54]. D. Ferrucci et A. Lally, 2004. UIMA: An architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering, 10(3-4): 327–348.

[55]. C. Friedman, G. Hripcsak, L. Shagina, et H. Liu, 1999. Arepresenting information in patient reports using natural language processing and the extensible markup language. Journal of the American Medical Informatics Association, 6: 76–87.

[56]. D. Demner-Fushman, W. W. Chapman, et C. J. McDonald, 2009. What can natural language processing do for clinical decision support? Journal of Biomedical Informatics, 42(5): 760–772.

[57]. BioNLP. http://www.bionlp.org/

[58]. ORBIT project. http://orbit.nlm.nih.gov/

[59]. Parikshit Sondhi, 2008. A Survey on Named Entity Extraction Techniques in the Biomedical Domain.

[60]. Ricardo A. Baeza-Yates et Berthier Ribeiro-Neto, 1999. Modern Information Retrieval, Addison-Wesley Longman Publishing Co., Inc., Boston, MA.

[61]. Christopher D Manning et Hinrich Schutze, 1999. Foundations of statistical natural language processing. MIT Press.

[62]. Anna Huang, 2008. Similarity measures for text document clustering. In Proceedings of the sixth new zealand computer science research student conference NZCSRSC2008, pages 49-56.

[63]. Pranam Kolari, Akshay Java, Tim Finin, Tim Oates, et Anupam Joshi, 2006. Detecting spam blogs: A machine learning approach. In Proceedings of the National Conference on Artificial Intelligence, pages 1351-1356. Association for the Advancement of Artificial Intelligence.

[64]. Lei Wu, Steven CH Hoi, et Nenghai Yu, 2010. Semantics-preserving bag-of-words models and applications. Image Processing, IEEE Transactions on, 19(7): 1908-1920.

[65]. Bridget T. McInnes et Ted Pedersen, 2013. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. Journal of Biomedical Informatics 46(6): 1116-1124.

[66]. Lynette Hirschman, Alexander A. Morgan, et Alexander S. Yeh, 2002. Rutabaga by anyother name: Extracting biological names. Journal of Biomedical Informatics.

[67]. FlyBase Consortium, 2002. The FlyBase database of the Drosophila genome projects and community literature. Nucleic Acids Res ;30(1):106–108.

[68]. Eason, Sergei Egorov, Anton Yuryev, et Nikolai Daraselia, 2004. A simple and practical dictionary-based approach for identification of proteins in medline abstracts. Journalof American Medical Information Association.

[69]. Sophia Ananiadou, 1994. A methodology for automatic term recognition. In Proceedings of COLING.

[70]. K. Fukuda, A. Tamura, T. Tsunoda, et T. Takagi, 1998. Toward information extraction: Identifying protein names from biological papers. In Proc. of Pacific Symposium on Bio-computing.

[71]. Denys Proux, Fransois Rechenmann, Laurent Julliard, Violaine Pillet, et Bernard Jacq, 1998. Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. In Proceedings of Ninth Workshop on Genome Informatics.

[72]. K. Frantzi, S. Ananiadou, et H. Mima, 2000. Automatic recognition of multi-word terms : the C-value/NC-value method,‖ in International Journal on Digital Libraries, vol. III, pp. 115–130.

[73]. A. Hliaoutakis, K. Zervanou, et E. G. M. Petrakis, 2009. The AMTExapproach in the medical document indexing and retrieval application‖, DataKnowledge Engineering, pp. 380–392.

[74]. F. Bechet, A. Nasr et F. Genet, 2000 "Tagging Unknown Proper Names Using Decision Trees", In proceedings of the 38th Annual Meeting of the Association for Computational Linguistics.

[75]. Nigel Collier, Chikashi Nobata, et Junichi Tsujii, 2000. Extracting the names of genes and gene products with a hidden markov model. In Proceedings of COLING.

[76]. McCallum A, Freitag D, et Pereira F., 2000 Maximum Entropy Markov Models for Information Extraction and Segmentation. Proc. ICML 2000.Stanford California.

[77]. Burr Settles, 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In Proceedings of JNLPBA.

[78]. Lawrence R. Rabiner, 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition, Proceedings of the IEEE, vol. 77, no 2, février 1989, p. 257–286 - DOI 10.1109/5.18626

[79]. Lafferty J, McCallum A et Pereira F, 2001, Conditional random fields: probabilistic models for segmenting and labeling sequence data. ICML' 2001.

[80]. Vladimir Vapnik and A. Lerner, Pattern Recognition using Generalized Portrait Method, Automation and Remote Control.

[81]. C. Cortes and V. Vapnik. Support-vector networks. Machine Learning, November 1995. 20:273–29.

[82]. Cover, T. M. et Thomas, J. A., 2006. Chapter 12, Maximum Entropy" (PDF). Elements of Information Theory (2 ed.). Wiley. ISBN 978-0471241959.

[83]. Stephen Della Pietra, Vincent Della Pietra, et John Lafferty, 1997. Inducing features of random fields. IEEE Transactions on P attern Analysis and Machine Intelligence. 19(4):380–393.

[84]. J. N. Darroch, et D. Ratcliff, 1972. Generalized iterative scaling for log-linear models. Annals of-Mathematical Statistics. 43(5):1470–1480

[85]. Rokach, Lior, et Maimon, O., 2008. Data mining with decision trees: theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711.

[86]. McCallum A, Freitag D, et Pereira, 2000. F: Maximum entropy Markov models for information extraction and segmentation. ICML' 02.

[87]. Wallach H, 2004. Conditional Random Fields: An Introduction. CIS, U of Pennsylvania2004.

[88]. Pinto D, McCallum A, Wei X, et Croft WB, 2003. Table extraction using conditional random fields. ACM SIGIR' 03.

[89]. Sha F, Pereira F, 2003. Shallow parsing with conditional random fields. HLT & NAACL' 03.

[90]. Zhu F, Shen B, 2012, Combined SVM-CRFs for Biological Named Entity Recognition with Maximal Bidirectional Squeezing. PLoS ONE 7(6): e39230. doi:10.1371/journal.pone.0039230

[91]. Tomoko Ohta, Yuka Tateisi, et Jin-Dong Kim, 2002. GENIA corpus: An annotated research abstract corpus in molecular biology domain. In Proceedings of Human Language Technology Conference.

[92]. Kim J, Ohta T, Tsuruoka Y, Tateisi Y, Collier N, 2004. Introduction to the bio-entity recognition task at JNLPBA. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. Stroudsburg, PA, USA: Association for Computational Linguistics:70–75.

[93]. Kim J, Ohta T, Tateisi Y, et Tsujii J, 2003: GENIA corpus-a semantically annotated corpus for bio-textmining. Bioinformatics2003,19:180–182

[94]. Cambridge Dictionaries Online, Cambridge University Press last access 2013, from http://dictionary.cambridge.org/dictionary/american -english/

[95]. Li, Y., Chung, S. M., et Holt, J. D, 2008. Text document clustering based on frequent word meaning sequences. Data Knowl. Eng., 64(1), 381 -404. doi: 10.1016/j.datak.2007.08.001

[96]. Hotho, A., Staab, S., et Stumme, G. ,2003. Text clustering based on background knowledge.

[97]. Miller, G. A.,1995. WordNet: a lexical database for English. Commun. ACM, 38(11), 39-41. doi: 10.1145/219717.219748

[98]. Wikipedia: About. From Wikipedia, the free encyclopedia., last access 2019, from http://en.wikipedia.org/wiki/Wikipedia:About

[99]. Bloehdorn, S., et Hotho, A, 2006. Boosting for text classification with semantic features . Proceedings of the 6th international conference on Knowledge Discovery on the Web: advances in Web Mining and Web Usage Analysis, Seattle, WA.

[100]. Ferretti, E., Errecalde, M., et Rosso, P., 2008. Does Semantic Information Help in the Text Categorization Task? Journal of Intelligent Systems, 17, 91 -107.

[101]. Bai, R., Wang, X., et Liao, J., 2010. Using an integrated ontology database to categorize web pages. Paper presented at the Proceedings of the 2010 international conference on Advances in computer science and information technology, Miyazaki, Japan.

[102]. Bloehdorn, S., et Hotho, A., 2006. Boosting for text classification with semantic features . Paper presented at the Proceedings of the 6th international conference on Knowledge Discovery on the Web: advances in Web Mining and Web Usage Analysis, Seattle, WA.

[103]. Aronson, A. R., et Lang, F. M., 2010. An overview of MetaMap: historical perspective and recent advances.

[104]. Zhong Z, et Ng H, 2010. It makes sense: a wide-coverage word sense disambiguation system for free text. In: Proceedings of the ACL 2010 system demonstrations, association for computational linguistics. p. 78–83.

[105]. Brody S, et Lapata M, 2009. Bayesian word sense induction. In: Proceedings of the 12th conference of the European chapter of the association for computational linguistics; p. 103–11.

[106]. Navigli R, Faralli S, Soroa A, de Lacalle O, et Agirre E, 2011. Two birds with one stone: learning semantic models for text categorization and word sense disambiguation. In: Proceedings of the 20th ACM international conference on Information and knowledge management. ACM; p. 2317–20

[107]. Kilgarri, Adam, 1998: Senseval: An exercise in evaluating word sense disambiguation programs. Proc. of the First International Conference on Language Resources and Evaluation.

[108]. Wessam Gad El Rab, Osmar R. Zaïane, et Mohammad El-Hajj, 2013: Biomedical text disambiguation using UMLS. ASONAM 2013: 943-947.

[109]. Alexopoulou D, Andreopoulos B, Dietze H, Doms A, Gandon F, Hakenberg J, et al., 2009, Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. BMC Bioinformatics; 10(1):28.

[110]. Jimeno-Yepes A, et Aronson A, 2010. Knowledge-based biomedical word sense disambiguation: comparison of approaches. BMC Bioinformatics; 11(1):569.

[111]. Garla VN, et Brandt C, 2012. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. J Am Med Inform Assoc 2012; 0(5):1–5.

[112]. Patwardhan S, Banerjee S, et Pedersen T, 2003. Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the fourth international conference on intelligent text processing and computational linguistics. p. 241–5

[113]. Rada R, Mili H, Bicknell E, et Blettner M, 1989. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 1989; 19(1):17–30.

[114]. Caviedes J, et Cimino J, 2004. Towards the development of a conceptual distance metric for the umls. J Biomed Inform 2004; 37(2):77–85.

[115]. Wu Z, et Palmer M, 1994. Verbs semantics and lexical selection. In: Proceedings of the 32nd meeting of association of computational linguistics. p. 133–8

[116]. Leacock C, et Chodorow M, 1998. Combining local context and WordNet similarity for word sense identification. WordNet: An Electron Lexical Database 1998; 49(2):265–83

[117]. Nguyen H, et Al-Mubaid H, 2006. New ontology-based semantic similarity measure for the biomedical domain. In: Proceedings of the IEEE international conference on granular computing. p. 623–8.

[118]. Resnik P, 1995. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th international joint conference on artificial intelligence. p. 448–53.

[119]. Jiang J, et Conrath D, 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings on international conference on research in computational linguistics. p. 19–33.

[120]. Lin D, 1998. An information-theoretic definition of similarity. In: Proceedings of the international conference on machine learning. p. 296–304. [http://citeseer.ist.psu.edu/95071.html].

[121]. Sánchez D, Batet M, et Isern D, 2011. Ontology-based information content computation. Knowledge-Based Syst; 24(2):297–303.

[122]. Lesk M, 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of the 5th annual international conference on systems documentation. p. 24–6

[123]. Banerjee S, et Pedersen T, 2003. Extended gloss overlaps as a measure of semantic relatedness. In: Proceedings of the 18th international joint conference on AI. p. 805–10.

[124]. Patwardhan S, et Pedersen T, 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In: Proceedings of the EACL 2006 workshop making sense of sense –bringing computational linguistics and psycholinguistics together, Trento, Italy. p. 1–8.

[125]. Jimeno-Yepes A, McInnes B, et Aronson A, 2011. An unsupervised vector approach to biomedical term disambiguation: integrating umls and medline. BMC Bioinformatics 2011; 12(1):223.

[126]. McInnes, B.T., Pedersen, T., Pakhomov, S.V.S., Liu, Y., et Melton-Meaux, G., 2006, UMLS: Similarity: Measuring the relatedness and similarity of biomedical concepts. In: HLT-NAACL, pp. 28–31

[127]. Sahami, M., et Heilman, T, 2006. "A web- based kernel function for measuring the similarity of short text snippets". In Proc. of WWW '06.

[128]. Pawlak, Z, 1991. Rough sets: Theoretical aspects of reasoning about data, Kluwer Dordrecht.

[129]. Ngo Chi Lang, 2003. A tolerance rough set approach to clustering web search results. Poland: Warsaw University.

[130]. Jan Komorowski, Zdzislaw Pawlak, 1998, L. P. A. S. Rough sets: a tutorial.

[131]. Adrzej Skowron, J. S, 1996. Tolerance approximation spaces. Fundamenta Informaticae 27, 2-3, 245-253.

[132]. Saori Kawasaki, et Ngoc Binh Nguyen, 2000, T. B. H. Hierarchical document clustering based on tolerance rough set model. In Principles of Data Mining and Knowledge Discovery, 4th European Conference, PKDD 2000, Lyon, France, September 13-16, 2000, Proceedings, D. A. Zighed, H. J. Komorowski, and J. M. Zytkow, Eds., vol. 1910 of Lecture Notes in Computer Science, Springer.

[133]. Tu Bao Ho, 2002, N. B. N. Nonhierarchical document clustering based on a tolerance rough set model. International Journal of Intelligent Systems 17, 2, 199 (212).

[134]. G. Salton , A. Wong , et C. S. Yang, 1975, A vector space model for automatic indexing, Communi-cations of the ACM, v.18 n.11, p.613-620, Nov.

[135]. Lund, K., Burgess, C., et Atchley, R. A., 1995. Semantic and associative priming in a high-dimensional semantic space. Cognitive Science Proceedings (LEA), 660-665.

[136]. Wen-tau Yih et Christopher Meek, 2007. Improving Similarity Measures for Short Segments of Text. Proceeding AAAI'07 Proceedings of the 22nd national conference on Artificial intelligence, V 2, pp. 1489-1494

[137]. Jones, R.; Rey, B.; Madani, O.; et Greiner, W. 2006. Generating query substitutions. Proc. of WWW conference

[138]. S. Deerwester, S. Dumais, G. W. Furnas, T. K. Landauer, et R. Harshman, 1990. Indexing by Latent Semantic Analysis. Journal of the Society for Information Science, vol. 41, no 6, pp 391-407

[139]. Thomas Hofmann, 1999. Probabilistic Latent Semantic Indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99)

[140]. D. M. Blei, Andrew Y. Ng, et Michael I. Jordan., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3. pp 993-1022

[141]. Adel, N., Crockett, K., Crispin, A., Chandran, D., et Carvalho, J. P., 2018. FUSE (Fuzzy Similarity Measure) - A measure for determining fuzzy short text similarity using Interval Type-2 fuzzy sets. 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE).

[142]. E. M. B. Nagoudi. D. Schwab, 2017. Semantic Similarity of Arabic Sentences with Word Embeddings. Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), Valencia, Spain, pp. 18–24

[143]. Nakamura, T., Shirakawa, M., Hara, T., et Nishio, S, 2014. Semantic Similarity Measurements for Multi-lingual Short Texts Using Wikipedia. 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT).

[144]. Shirakawa, M., Nakayama, K., Hara, T., et Nishio, S., 2013. Probabilistic semantic similarity measurements for noisy short texts using Wikipedia entities. Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management - CIKM

[145]. Sahami, M., et Heilman, T, 2006. "A web-based kernel function for measuring the similarity of short text snippets". In Proc. of WWW 2006.

[146]. Jiliang TANG, Xufei WANG, Huiji GAO, Xia HU et Huan LIU, 2012. Enriching short text representation in microblog for clustering Front. Comput. Sci., 6(1) DOI 10.1007/s11704- 009- 0000- 0.

[147]. Jimeno-Yepes A, McInnes B, Aronson A, 2011. An unsupervised vector approach to biomedical term disambiguation: integrating umls and medline. BMC Bioinformatics 2011; 12(1) :223.

[148]. Wessam Gad El Rab, Osmar R. Zaïane, et Mohammad El-Hajj, 2013: Biomedical text disambiguation using UMLS. ASONAM 2013: 943-947.

[149]. Zakaria Elberrichi, Malika Taibi, et Amel Belaggoun, 2012. "Multilingual Medical Documents Classification Based on MesH Domain Ontology". CoRR abs/1206.4883.

[150]. A. Amine, Z. Elberrichi, et M. Simonet, 2010. Evaluation of Text Clustering Methods Using WordNet, International Arab Journal of Information Technology, Vol. 7, 2010, pp. 351.

[151]. Guyot J., Radhoum S., et Falquet.G, 2005. Ontology-Based Multilingual Information Retrieval. In CLEF, 2005.

[152]. Litvak M., Last M., et Kisilevich S., 2006. Improving classification of multi-lingual web documents using domain ontologies, in KDO05, The Second International Workshop on Knowledge Discovery and Ontologies. Porto, Portugal October 7th. 2006.

[153]. Sanchez D., et Moreno A., 2004 "Creating ontologies from Web documents". Recent Advances in Artificial Intelligence Research and Development. IOS Press, vol. 113, pp.11-18.

[154]. M. H. Song, S. Y. Lim, S. B. Park, et D. J. Kang, 2005. An automatic approach to classify web documents using a domain ontology. In The first international conference on pattern recognition and machine intelligence Kolkata, India 2005, pp. 666-671.

[155]. Séaghdha, D.O¸ 2009: Semantic classification with WordNet kernels. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp. 237–240. Association for Computational Linguistics, Boulder.

[156]. Lan, M., et al, 2009. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. IEEE Trans. Pattern Anal. Mach. Intell. 31(4), 721–735.

[157]. Li, Z., Li, P., Wei, W., Liu, H., He, J., Liu, T., et Du, X., 2009. AutoPCS: A Phrase-Based Text Categorization System for Similar Texts. In: Li, Q., Feng, L., Pei, J., Wang, S.X., Zhou, X., Zhu, Q.-M., et al. (eds.) APWeb/WAIM 2009. LNCS, vol. 5446, pp. 369–380. Springer, Heidelberg.

[158]. Bloehdorn, S., et Hotho, A., 2006. Boosting for Text Classification with Semantic Features. In: Mobasher, B., Nasraoui, O., Liu, B., Masand, B. (eds.) WebKDD 2004. LNCS (LNAI), vol. 3932, pp. 149–166. Springer, Heidelberg.

[159]. Bai, R., Wang, X., et Liao, J., 2010. Using an Integrated Ontology Database to Categorize Web Pages. In: Kim, T.-H., Adeli, H. (eds.) AST/UCMA/ISA/ACN 2010. LNCS, vol. 6059, pp. 300–309. Springer, Heidelberg.

[160]. Guisse, A., Khelif, K., et Collard, M., 2009. PatClust: une plateforme pour la classification sémantique des brevets. In: Conférence d'Ingénierie des connaissances, Hammamet, Tunisie.

[161]. Peng, X., et Choi, B. 2005. Document classifications based on word semantic hierarchies. In: International Conference on Artificial Intelligence and Applications (AIA 2005), pp. 362–367.

[162]. Wang, J.Z., et Taylor, W., 2006: Concept Forest: A New Ontology-assisted Text Document Similarity Measurement Method. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence 2007, pp. 395–401. IEEE Computer Society.

[163]. Stein, B., Eissen, S.M.Z., et Potthast, M., 2006. Syntax versus semantics: Analysis of enriched vector space models. In: Third International Workshop on Text-Based Information Retrieval (TIR 2006). University of Trento, Italy.

[164]. Hersh, W., et al., 1994. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 192–201. New York, Inc., Dublin.

[165]. Sokolova, M., et Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing Management 45(4), 427–437.

[166]. Sanderson, M, 2000. Retrieving with Good Sense. In Information Retrieval, Vol. 2(1), Pp 49 – 69.

[167]. Stokoe, M. P. Oakes, et J. Tai, 2003. Word sense disambiguation in information retrieval revisited. In Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 159–166.

[168]. S. B. Kim, H. C. Seo, et H. C. Rim., 2004. Information retrieval using word senses: root sense tagging approach. In Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 258–265

[169]. H. Fang., 2008. A re-examination of query expansion using lexical resources. In Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies, pages 139–147.

[170]. E. Agirre, X. Arregi, et A. Otegi., 2010. Document expansion based on WordNet for robust IR. In Proceedings of the 23rd International Conference on Computational Linguistics, pages 9-17.

[171]. Jihen Majdoubi, Hatem Loukil, Mohamed Tmar, et Faïez Gargouri, 2012: An approach based on language modeling for improving biomedical information retrieval. International Journal of Knowledge-based and Intelligent Engineering Systems, vol. 16, no. 4, pp. 235-246.

[172]. D. Dinh, et L. Tamine, 2010. Sense-based biomedical indexing, and retrieval, in International Conference on Applications of Natural Language to Information Systems, Springer-Verlag, Cardiff, UK, pp. 24–35.

[173]. Albitar S., Fournier S., et Espinasse B., 2012. The Impact of Conceptualization on Text Classification. In: Wang X.S., Cruz I., Delis A., Huang G. (eds) Web Information Systems Engineering - WISE 2012. WISE 2012. Lecture Notes in Computer Science, vol 7651. Springer, Berlin, Heidelberg

[174]. Mohammed Rais et Abdelmonaime Lachkar, 2016: Evaluation of Disambiguation Strategies on Biomedical Text Categorization. LNBI – Springer Verlag IWBBIO 2016: 790-801

[175]. Mohammed Rais et Abdelmonaime Lachkar, 2016: Biomedical Word Sense Disambiguation context-based: Improvement of SenseRelate method. IEEE Explore - 2016 International Conference on Information Technology for Organizations Development (IT4OD)

[176]. Michael Dittenbach, 2010. Scoring and Ranking Techniques - tf-idf term Weighting and Cosine Similarity. http://www.ir-facility.org/scoring-and-ranking-techniques-tf-idf-term-weighting-and-cosine-similarity.

[177]. "What does tf-idf mean? How to Compute" Information Retrieval and Text Mining http://www.tfidf.com/

[178]. Voorhees, E.M., et Harman, D.K., 2005. TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge.

[179]. Baziz, M., 2005. Indexation conceptuelle guidée par ontologie pour la recherche d'information. Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

[180]. Krovetz, R. et Croft, W. B., 1992. Lexical ambiguity and information retrieval. ACM Transactions on Information Systems, 10(2):115–141.

[181]. Voorhees, E. M., 1993. Using WordNet to Disambiguate Word Senses for Text Retrieval. In SIGIR, pages 171–180.

[182]. Sanderson, M., 1994. Word sense disambiguation and information retrieval. In SIGIR, pages 142–151.

[183]. Mihalcea, R., Tarau, P. et Figa, E., 2004. PageRank on semantic networks with application to word sense disambiguation. In International Conference on Computational Linguistics (COLING), pages 1126–1132

[184]. Liu, S., Liu, F., Yu, C. et Meng, W., 2004. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'04, pages 266–272, New York, NY, USA. ACM

[185]. Boubekeur, F., 2008. Contribution à la définition de modèles flexibles de recherche d'information basés sur les CP-Nets. Thèse de doctorat, Université Paul Sabatier.

[186]. Ounis, I., Amati, G., V., P., He, B., Macdonald, C., et Johnson 2005. Terrier Information Retrieval Platform. In Proceedings of the 27th European Conference on IR Research (ECIR 2005), volume 3408 de Lecture Notes in Computer Science, pages 517–519. Springer

[187]. Yongjing Lin, Wenyuan Li, Keke Chen, et Ying Liu, 2007. Model Formulation: A Document Clustering and Ranking System for Exploring MEDLINE Citations. JAMIA 14(5): 651-661

[188]. Müller HM, Kenny EE, Sternberg PW (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol. 2(11) e309. PMID: 15383839.

[189]. Perez-Iratxeta C, Pérez AJ, Bork P, et Andrade MA, 2003. Update on XplorMed: A web server for exploring scientific literature. Nucleic Acids Res. 31(13) 3866-3868. PMID: 12824439

[190]. Doms A, et Schroeder M, 2005. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res. 33 (Web Server issue) W783-6. PMID: 15980585

[191]. Allende RA, 2009. Accelerating searches of research grants and scientific literature with novo|seek. Nature Methods volume6, page394.

[192]. Wermter J, Tomanek K, et Hahn U, 2009. High-performance gene name normalization with GeNo. Bioinformatics. 25(6) 815-821. PMID: 19188193.

[193]. Nelson SJ et al, 2001. Relationships in Medical Subject Headings (MeSH) in Relationships in the organization Of knowledge Kluwer Acdamic Publisher, Netherlands, 171: 184.

[194]. Olivier Bodenreider, 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32 (Database issue): D267-70. PMID: 14681409

[195]. Wille, R., 2005. Formal concept analysis as mathematical theory of concepts and concept hierarchies. Form. Concept Anal. 1–33.

[196]. Carpineto, C., Romano, G., 2004. Exploiting the potential of concept lattices for information retrieval with CREDO. J. Univers. Comput. Sci. 10, 985–1013

[197]. Cigarrán, J.M., Gonzalo, J., Peñas, A., Verdejo, F., (2004). Browsing search results via formal concept analysis: automatic selection of at-tributes. Concept Lattices 2961, 201–202

[198]. Zhang, Y., et Feng, B., 2008. Clustering search results based on formal concept analysis. Inf. Technol. J.

[199]. Carpineto, C., Osin´ski, S., Romano, G., et Weiss, D., 2009. A survey of Web clustering engines. ACM Comput. Surv. 41, 1–38.

[200]. Kaufman, L., et Rousseeuw, P. J., 2005. Finding groups in data: an introduction to cluster analysis. Intensive Care Med. 33, 368.

[201]. Hartigan, J.A., et Wong, M.A., 1979. A K-means clustering algorithm. J. R. Stat. Soc. 28, 100–108.

[202]. Planck, M., et Luxburg, U. Von, 2006. A tutorial on spectral clustering a tutorial on spectral clustering. Stat. Comput. 17, 395–416.

[203]. Maarek, Y.S., Fagin, R., Ben-shaul, I.Z., et Pelleg, D., 2000. Ephemeral document clustering for web applications. IBM Res. Rep. RJ 10186, 1–26

[204]. Lawrie, D.J., et Croft, W.B., 2003. Generating hierarchical summaries for web searches. In: Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Dev. informaion Retr. –SIGIR ’03 457

[205]. Geraci, F., Pellegrini, M., Maggini, M., et Sebastiani, F., 2006. Cluster generation and cluster labelling for Web snippets: a fast and accurate hierarchical solution. String Process. Inf. Retr. 13, 25–36

[206]. Cutting, D.R., 1992. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections 1 Introduction 2 Scat-ter/Gather Browsing.

[207]. Zamir, O., et Etzioni, O., 1999. Grouper: A dynamic clustering interface to Web search results. In: Proc. WWW8.

[208]. Ch, A.K., Dias, S.M., et Vieira, N.J., 2015. Knowledge reduction in formal contexts using non-negative matrix factorization. Math. Com-put. Simul. 109, 46–63.

[209]. Cheung, K.S.K., et Vogel, D., 2005. Complexity reduction in lattice-based information retrieval. Inf. Retr. Boston. 8, 285–299.

[210]. Dias, S.M., et Vieira, N., 2010. Reducing the size of concept lattices: the JBOS approach. In: Cla 2010, pp. 80–91.

[211]. Dias, S.M., et Vieira, N.J., 2015. Concept lattices reduction: definition, analysis and classification. Expert Syst. Appl. 42, 7084–7097.

[212]. Li, J., Mei, C., et Lv, Y., 2012. Knowledge reduction in real decision formal contexts. Inf. Sci. (Ny) 189, 191–207

[213]. Smalheiser, N.R., Zhou, W., et Torvik, V.I., 2008. Anne O'Tate: a tool to support user-driven summarization, drill-down and brows-ing of PubMed search results.J. Biomed. Discov. Collab., 3,2.

[214]. Yamamoto, Y. et Takagi, T., 2007. Biomedical knowledge navigation by literature clustering. J. Biomed. Inform., 40, 114–130

[215]. Perez-Iratxeta, C., 2001. XplorMed: a tool for exploring MEDLINE abstracts. Trends Biochem. Sci., 26, 573–575

[216]. Ganter, B., 2003. Ch1 & Ch2: Contexts, concepts, and concept lattices. Form. Concept Anal. Methods Appl. Comput. Sci.

[217]. Godin, R., Missaoui, R., et Alaoui, H., 1995. Incremental concept formation algorithms based on Galois (concept) lattices. Comput. Intell. 11, 246–267

[218]. Bordat, J., 1986. Calcul pratique du treillis de Galois d'une correspondance. Math. Sci. Hum. Math. Soc. Sci. 96, 31–47

[219]. Nourine, L., et Raynaud, O., 2002. A fast incremental algorithm for building lattices. J. Exp. Theor. Artif. Intell. 14, 217–227.

[220]. William R. Hersh, Aaron M. Cohen, Lynn Ruslen, et Phoebe M. Roberts, 2007. TREC 2007 Genomics Track Overview. TREC 2007.

[221]. Beitzel S.M., Jensen E.C., et Frieder O., 2009. MAP. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.

[222]. Craswell N., 2009. Bpref. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.