Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École doctorale Mathématiques, sciences de l'information et de l'ingénieur | ED 269 Université de Strasbourg

ÉCOLE DOCTORALE MATHEMATIQUES, SCIENCES DE L'INFORMATION ET DE L'INGENIEUR – ED269

UMR7357

THÈSE présentée par :

Karim EL HAFF

soutenue le : **05 novembre 2024**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Informatique

Extraction et modélisation d'informations textuelles pour l'exploitation de pharmacopées arabes anciennes

THÈSE dirigée par :

Mme LE BER Florence	Prof., université de Strasbourg
Mme PITCHON Véronique	Prof., université de Strasbourg

RAPPORTEURS :

M. DOUCET Antoine	Prof., université de La Rochelle
M. ROCHE Mathieu	Dr. HDR, CIRAD Montpellier

AUTRES MEMBRES DU JURY :

Mme BRAUD Agnès	As. Prof., université de Strasbourg
M. BURRI Sylvain	Dr., université Toulouse Jean Jaurès

INVITÉS :

M. FECHTER Pierre	Prof., université de Strasbourg

كعيون للكون نحن ، في أعيننا ينعكس ، العقل مركبنا في بحر الأزل

As the eyes of the universe we, and in them it is reflected,
the mind is our ship sailing the sea of the infinite..

وبنا يستكشف الكون ذاته ، في لحن الحوار بين الأرض والفضاء

..and the universe explores itself through us,
the conscious observer, along the melody of
dialogue between the earth and space

Acknowledgements

At the end of these three years of research, it is important for me that I take a moment to express my sincere gratitude to those who played a decisive role in bringing this work to fruition. This manuscript, the fruit of a complex and rewarding journey, is the result of unfailing collaboration and support, and I would like to pay tribute to all those who contributed to the success of this scientific adventure.

First and foremost, I would like to express my sincere gratitude to my co-advisors Prof. Florence Le Ber and Prof. Véronique Pitchon, as well as my supervisor Prof. Agnes Braud, without whom this dissertation would not have been possible. Under their advisorship and supervision, I enjoyed two key elements that I never imagined having as a graduate student: unlimited academic freedom, but simultaneously, absolute academic care. Indeed, I thank them for their patience, motivation, frequent productive meetings, and immense transmission of knowledge throughout this project; Prof. Florence Le Ber and Prof. Agnes Braud for their help in all matters related to the computational perspective and Prof. Véronique Pitchon for her help regarding the historical side of things.

Beyond my advisors, I would like to thank the rest of my thesis committee: Prof. Antoine Doucet, Prof. Mathieu Roche, and Prof. Sylvain Burri, for partaking in this step of my academic endeavor, and I am highly looking forward to their comments, questions and constructive critical insights during the defense that will allow me to improve my perspective and knowledge going forward.

I would like to sincerely thank everyone from the MANSA team who has collaborated with me for the advancement of this work. Prof. Pierre Fechter and Prof. Régine Janel-Bintz for their knowledge and help in all matters related to biology, Prof. Catherine Vonthron-Sénécheau and Sergio Enrique Ortiz Aguirre for all their support in all matters related to botany and pharmacognosy, Dr. Elhoussaine Oussialy for his vast knowledge and help in the historical perspective and annotations, alongside Ing. Mohammed Benkhalid in all matters related to the Digital Humanities. The MANSA group is a big family and ever so growing, and I would also like to thank everyone that is part of the team and actively working towards uncovering new knowledge from ancient wisdom.

I would like to express my gratitude to aspiring botanist Anthony Massiala, who helped me construct the plant-based database, as well as PhD Candidate of Pharmacognosy Capucine Braillon who immensely helped in the annotation process and provided me with regular updates on the lab-experiment side of things.

My gratitude also goes to the people with whom I have collaborated alongside my advisorship and supervision in producing scientific publications; PhD Candidate in NLP Wissam Antoun with whom I have collaborated and produced two publications throughout my PhD, as well as PhD Candidate in Computer Science Vanessa Fokou and Prof. Xavier Dolques, who both immensely helped me in the field of Formal and Relational Concept Analysis which was a new domain for me, and I have grown more familiar and at ease with such methods thanks to their help.

I also thank Prof. Nicholas Lachiche, Prof. Pierre Martin and Prof. Loup Bernard who were part of my thesis monitoring committee for two consecutive years and whose interesting comments and insightful critiques helped me pave the way for an improved way of working in an academic setting and helped fine-tune the flow of my progress.

I am also thankful for everyone at the SDC team at ICube and everyone at MISHA for their kindness and hospitality during my years here.

My gratitude also extends to the Professors from my former Master's degree, "Technologie des Langues", Prof. Delphine Bernhard, Prof. Pablo Ruiz Fabo and Prof. Amalia Todirascu who first introduced me to the world of NLP in the first place and shaped my knowledge in it, and with whom I have been still in contact during my PhD, through their encouragement of me revisiting the faculty, but this time through the lens of a teacher. Indeed, I thank them for offering me the opportunity to teach classes that I formerly took with them as a student, this made me discover a passion for pedagogy and the transmission of knowledge.

I would like to acknowledge and thank my friends who have made this journey memorable. To those from different horizons with whom I had countless thought-provoking discussions, engaged in cultural activities, traveled, played music, and even, funnily enough, some of whom I collaborated on designing a video game inspired by my thesis in the context of a scientific game jam (see Appendix A: Maison de la Sagesse, a game inspired by this thesis)— thank you for your companionship and creativity. Each of you contributed uniquely to my experience, and I appreciate the moments we shared.

Finally, I would like to express my deepest gratitude to my family. To my parents, Samir, Haifa, my siblings, Sarah, Zeina and Abdallah and their families, for their unconditional love and support throughout my academic endeavors. Their encouragement has been a constant source of strength, even in the face of the country-wide difficulties and tragedies unfolding back home. Despite the challenges and uncertainties, my family has never wavered in their support, offering both moral and practical assistance whenever needed. Their resilience and commitment to my progress have been inspirational, and for that, I am grateful. This achievement is as much theirs as it is mine.

Publications

Available:

- K. El Haff, A. Braud, F. Le Ber, et V. Pitchon, « Modélisation des ingrédients de remèdes issus de pharmacopées arabes médiévales dans une base de données graphe », Actes 34es Journ. Francoph. D'Ingénierie Connaiss. Plate-Forme Intell. Artif. PFIA 2023 Jul 2023 Strasbg. Fr., p. 34, 2023, https://inria.hal.science/hal-04162861v1
- K. El Haff, W. Antoun, F. Le Ber, et V. Pitchon, « Reconnaissance des entités nommées pour l'analyse des pharmacopées médiévales », in EGC 2023 - Extraction et Gestion des Connaissances, Lyon, France, 2023. [En ligne]. Disponible sur: https://editions-rnti.fr/?inprocid=1002834
- V. Fokou, K. El Haff, A. Braud, X. Dolques, F. Le Ber, V. Pitchon, "Exploring Old Arabic Remedies with Formal and Relational Concept Analysis" in CONCEPT, Cádiz, 2024, https://link.springer.com/chapter/10.1007/978-3-031-67868-4_20

Accepted:

- K. El Haff, W. Antoun, A. Braud, F. Le Ber, et V. Pitchon, « Building and Assessing a Named Entity Recognition Resource for Ancient Pharmacopeias », in 27TH European Conference On Artificial Intelligence, Santiago de Compostela, October 2024

In addition to the topics studied in this dissertation which are mentioned above, I worked during my master's degree on the creation of the first morphologically annotated corpus for my native Lebanese Levantine Arabic dialect, published at LREC 2022.

- K. El Haff, M. Jarrar, T. Hammouda, and F. Zaraket, « Curras + Baladi: Towards a Levantine Corpus. » In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 769–778, Marseille, France. European Language Resources Association, 2022, https://aclanthology.org/2022.lrec-1.82/

Résumé en français

Introduction

Cette thèse s'inscrit dans un contexte d'analyse de textes anciens de pharmacopées arabes par des méthodes de traitement automatique du langage naturel. Elle a comme but principal la recherche de connaissances intéressantes pour la conception de nouveaux médicaments et de potentiels alternatifs aux antibiotiques. Pour ce faire, plusieurs étapes sont mises en œuvre. En effet, ce type de travail consiste à manipuler une grande quantité de données textuelles afin de les représenter sous la forme d'une base de données de type graphe et ensuite développer un système de questionnement qui permettra la constatation de nouvelles connaissances.

Les principaux textes sur lesquels nous travaillons ont été traduits en anglais, ce sont les ouvrages suivants :

- *Sābūr ibn Sahl's Dispensatory in the Recension of the ʿAḍudī Hospital (9th century Baghdad)*, traduit par Oliver Kahl[1].
- *The Dispensatory of Ibn at-Tilmīd (12th century Baghdad)*, traduit par Oliver Kahl[2].
- *Ibn al-Jazzar's Provision for the Traveler and Nourishment for the Sedentary (10th century, Kairouan) - Book 7,* traduit par Gerrit Bos[3].

La problématique principale posée est la suivante : Comment passer automatiquement de données textuelles historiques non-structurées en langage naturel à des données structurées et exploitables ?

L'objectif principal de la thèse est donc de mettre en œuvre et d'adapter un ensemble d'approches, permettant de définir une chaîne automatique de traitements s'appliquant aux textes pour en extraire les données utiles, les structurer, pour ensuite les interroger. Ce travail de traitement automatique du langage naturel s'inscrit dans le domaine du Text Mining, se focalisant sur les différentes tâches de la reconnaissance des entités nommées, du liage des entités nommées, ainsi que la modélisation des informations extraites.

---

[1] O. Kahl, Sābūr Ibn Sahl's Dispensatory in the Recension of the `Aḍudī Hospital. BRILL, 2009.

[2] O. Kahl, The Dispensatory of Ibn at-Tilmīḏ: Arabic Text, English Translation, Study and Glossaries. BRILL, 2007.

[3] G. Bos, Ibn al-Jazzar's Zad al-musafir wa-qut al-hadir, Provision for the Traveller and Nourishment for the Sedentary, Book 7 (7–30) Critical Edition of the Arabic Text with English Translation, and Critical Edition of Moses ibn Tibbon's Hebrew Translation (Ṣedat ha-Derakhim). 2015

Contexte général

Parmi les travaux importants d'exploration manuelle de pharmacopées anciennes, il y a celui de Harrison et al.[4] qui estime que les sociétés médiévales utilisaient une série de substances naturelles pour traiter des symptômes identifiables aujourd'hui comme des infections microbiennes ; ceci serait donc reflété dans leurs pharmacopées. Ils ont identifié et reconstitué un remède potentiel pour l'infection par *Staphylococcus Aureus* à partir d'un livre médical anglo-saxon du Xe siècle. Le remède a tué à plusieurs reprises les bactéries de biofilms dans un modèle in vitro d'infection des tissus mous. Il a également détruit le *Staphylococcus Aureus* résistant à la méticilline dans un modèle de plaie chronique chez une souris. L'efficacité du remède est liée à l'action combinée de plusieurs de ses ingrédients, ce qui démontre le potentiel des anciennes pharmacopées comme source de connaissances médicales. Plus récemment, le travail d'analyse des données d'une pharmacopée médiévale britannique, avec une saisie manuelle des données, a permis de découvrir des tendances intéressantes dans les corpus explorés[5], ce qui confirme l'intérêt de passer d'une extraction manuelle des données à une automatisation de l'extraction des données.

En effet, l'automatisation de l'exploration de données textuelles permet d'extraire des informations plus rapidement à partir de grands volumes de données non structurées, souvent difficiles à traiter manuellement. Les techniques de fouille de textes sont de plus en plus utilisées dans des domaines variés dont la médecine. Les remèdes des manuscrits que nous traitons sont décrits principalement par des noms de plantes et d'ingrédients à base animale ou minérale, les symptômes sont dénommés par des appellations issues du langage courant et non par les appellations scientifiques standardisées. Bien que des travaux sur la médecine moderne soient représentés dans les projets de fouille de textes[6], nous constatons un manque de jeux de

---

[4] F. Harrison, A. E. L. Roberts, R. Gabrilska, K. P. Rumbaugh, C. Lee, et S. P. Diggle, « A 1,000-Year-Old Antimicrobial Remedy with Antistaphylococcal Activity », mBio, vol. 6, no 4, p. e01129-15, août 2015, doi: 10.1128/mBio.01129-15.

[5] E. Connelly, C. I. del Genio, et F. Harrison, « Data Mining a Medieval Medical Text Reveals Patterns in Ingredient Choice That Reflect Biological Activity against Infectious Agents », mBio, vol. 11, no 1, p. e03136-19, févr. 2020, doi: 10.1128/mBio.03136-19.

[6] B. Wang, Q. Xie, J. Pei, P. Tiwari, Z. Li, et J. Fu, « Pre-trained Language Models in Biomedical Domain: A Systematic Survey », CoRR, vol. abs/2110.05006, 2021, [En ligne]. Disponible sur: https://arxiv.org/abs/2110.05006

données provenant de pharmacopées anciennes. Ainsi, il serait intéressant d'explorer cette voie technologique pour ce qui relève de ces connaissances anciennes.

Reconnaissance des entités nommées

Le premier grand axe de travail de ce projet de recherche se concentre sur l'application de la reconnaissance d'entités nommées (NER) aux traductions anglaises de pharmacopées de l'époque abbaside. En effet, c'est à partir de la tâche NER que peuvent se déclencher les autres tâches nécessaires pour aboutir à la finalité du de la fouille de textes.

En premier lieu, un travail ayant comme but de trouver le modèle Transformer le plus performant pour la tâche a été fait. Pour entraîner le modèle, nous avons effectué l'annotation des remèdes d'une pharmacopée en intégralité, la traduction anglaise de l'ouvrage « Sābūr ibn Sahl's Dispensatory in the Recension of the ʿAḍudī Hospital » par Oliver Kahl qui compte 36 961 tokens divisés en 292 remèdes. Ce document a été annoté manuellement pendant un mois sous la forme d'un tableur de type CSV.

Pour effectuer l'annotation, 4 types d'étiquettes ont été utilisés :

- ***Type*** : forme du remède (pastille, pilule, etc.)
- ***Sym*** : symptôme de maladie
- ***Ing*** : ingrédient utilisé
- ***Org*** : organe

Les données ont été annotées dans le format IOB2 (abréviation de « inside, outside, beginning »)

Suite à notre étude qui constate que DeBERTaV3[7] obtient le meilleur résultat, le but était d'élargir le corpus pour développer un modèle de NER capable d'identifier et de classer avec plus de précision des entités telles que les ingrédients, les symptômes, les organes et les types de préparations à partir de ces textes. De plus, cet élargissement du corpus permet d'analyser les résultats qualitativement et quantitativement, ayant plusieurs sources pour les corpus annotés. Afin d'analyser les résultats :

- Nous testons la transférabilité entre les traducteurs, en entraînant notre modèle sur les manuscrits traduits par Oliver Kahl, à savoir ceux d'Ibn Tilmīdh et de Sabur Ibn Sahl, puis en les testant sur le manuscrit d'Ibn Jazzar traduit par Gerrit Bos, et vice-versa.
- Nous examinons l'effet de la transférabilité entre les auteurs originaux.
- Nous étudions l'effet de la combinaison des traducteurs des manuscrits d'entraînement.

[7] Developpé par P. He, X. Liu, J. Gao, et W. Chen, « DeBERTa: Decoding-enhanced BERT with Disentangled Attention ». arXiv, 6 octobre 2021. doi: 10.48550/arXiv.2006.03654.

- Nous examinons également l'effet de la taille de l'ensemble de données d'entraînement.

La collaboration entre linguiste-informaticien, pharmacognosistes et historiens a abouti à un corpus annoté manuellement comprenant environ 92 000 tokens (16K pour Ibn Jazzar, 40K pour Ibn Tilmīdh et 36K pour Sabur Ibn Sahl), chaque token étant étiqueté selon sa pertinence contextuelle suivant la méthode IOB2.

Pour atteindre l'objectif de tester la transférabilité vers un nouveau traducteur, nous avons catégorisé nos expériences en plusieurs groupes distincts afin d'évaluer systématiquement les performances de notre modèle de NER.

L'impact de la variance des traducteurs sur la performance du modèle est tangible, ce qui indique une généralisabilité du modèle. En entraînant sur des manuscrits traduits par Oliver Kahl (Ibn Tilmīdh et Sabur Ibn Sahl) et en testant sur la traduction de Gerrit Bos (Ibn Jazzar), des scores F1 de 69,09 (Ibn Tilmīdh) et 69,57 (Sabur) ont été atteints, démontrant une baisse notable mais pas invalidante par rapport au score de test intra-traducteur de 75,63 (Ibn Jazzar entraîné et testé). Lorsque la taille complète du corpus est utilisée pour un entraînement avec un ensemble de validation croisée à 5 splits sans aucun remaniement, le modèle obtient un score F1 de 89,7% sur le meilleur Fold.

Les résultats montrent que l'entraînement uniquement sur des manuscrits par Ibn Tilmīdh ou Sabur Ibn Sahl donne des scores similaires sur leurs ensembles de test respectifs, suggérant que la performance du modèle n'est pas affectée par l'auteur du manuscrit original, contrairement au changement de traducteurs. Mélanger les données de différents manuscrits atténue clairement l'effet du traducteur et augmente légèrement la performance globale. L'entraînement sur une combinaison d'Ibn Jazzar et d'Ibn Tilmīdh ou d'Ibn Jazzar et de Sabur Ibn Sahl a donné des scores plus élevés sur des manuscrits d'un traducteur différent, par rapport à l'entraînement uniquement sur un seul traducteur, indiquant que la diversification des données d'entraînement, même à travers les traducteurs, contribue positivement à la robustesse du modèle. En comparant les stratégies d'élargissement des données à partir d'une source unique et d'incorporation de traductions provenant de divers auteurs ou traducteurs, nos résultats favorisent la diversification.

Liage des entités nommées

Le deuxième axe principal du travail vise à implémenter et affiner un modèle de désambiguïsation et de liage d'entités nommées (Named Entity Linking/NEL) afin de répondre au besoin d'obtenir des informations scientifiques automatiquement à partir du nom courant d'une plante dans le texte, retrouvée automatiquement par le modèle NER.

Préparation des données : Le travail a commencé par l'intégration des données textuelles à l'issue de l'annotation automatique du modèle NER, transformées en un format exploitable, notamment un tableur CSV. Un script Python a été développé pour fusionner les entités nommées successives constituées de plusieurs mots et pour simplifier les balises associées, en éliminant les préfixes (comme 'B-' et 'I-') et en gérant les valeurs manquantes. Cette étape a été nécessaire pour assurer un corpus cohérent et prêt pour l'analyse sémantique et la désambiguïsation.

Traitement et désambiguïsation : Le cœur de ce travail consiste en l'application d'un algorithme de désambiguïsation, utilisant la ressource BabelNet, qui est un dictionnaire et un réseau sémantique multilingue[8], pour relier les termes extraits à des concepts univoques. Les phases de traitement incluent la lemmatisation, la suppression des *stop-words* et l'identification des parties des plantes et de leurs transformations à partir de dictionnaires pour réduire le bruit dans l'identification des entités. Chaque terme retrouvé dans BabelNet est ensuite passé par un processus de liage pour associer les mentions de plantes à des identifiants uniques dans une bases de connaissances globale comme Wikidata qui permettrait l'identification de l'identifiant unique qui se trouve dans la base Global Biodiversity Information Facility (GBIF)[9].

Analyse de la distribution géographique : Une composante innovante du travail est l'analyse de la distribution géographique des plantes mentionnées dans les textes pour les désambiguïser. En effet, la base GBIF contient des géo-références pour chaque taxon identifié. Pour chaque taxon, il peut y avoir plusieurs espèces ou sous-espèces différentes au sein de la distribution

---

[8] R. Navigli, M. Bevilacqua, S. Conia, D. Montagnini, et F. Cecconi, « Ten Years of BabelNet: A Survey », in Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization, août 2021, p. 4559-4567. doi: 10.24963/ijcai.2021/620.

[9] GBIF - https://www.gbif.org/

géographique, ce qui peut créer davantage d'ambiguïté par rapport à la plante réellement utilisée par le médecin abbaside.

Afin de désambiguïser la plante, un algorithme qui prend en compte l'origine géographique du manuscrit analysé a été développé. En effet, nous estimons que si une pharmacopée a été rédigée à Bagdad, les plantes décrites dans cette pharmacopée sont plus probablement référencées dans un rayon autour de la région de Bagdad. Ayant accès aux coordonnées des géo-références d'un taxon contenu dans GBIF, nous calculons de la densité des références dans différentes zones géographiques définies autour d'un point choisi. Un système de notation a été créé pour classer les différents candidats en fonction de la densité (pour les rayons de 50, 500 et 1500km).

À l'issue de ce classement avec les notations assignées, nous pouvons lier la plante mentionnée dans le texte au meilleur candidat situé dans la base GBIF. Ensuite, grâce aux informations disponibles pour les taxons recensés dans GBIF, nous pouvons extraire les détails importants, notamment le nom scientifique et la famille de la plante.

Le taux de réussite global est de 32,44 %. Ce taux faible illustre les difficultés rencontrées pour maintenir un niveau élevé de rétention des entités à plusieurs stades du traitement.

Modélisation des informations

À l'issue des deux premières tâches concernant la reconnaissance et le liage des entités nommées importantes des manuscrits, les données peuvent être structurées formellement pour être stockées dans une base de données (BDD) orientée graphe qui est adaptée pour les données qui sont à la fois qualitatives et relationnelles. En effet, ceci est directement possible après un post-traitement du tableur d'annotation enrichi. Les données sont ensuite stockées dans la BDD selon le modèle illustré par la figure 1.

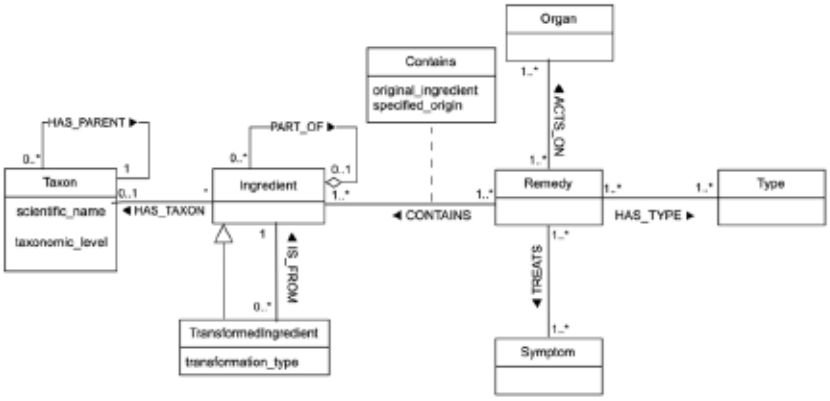


Figure 1 - Modèle UML de la base de données orientée graphe

Nous nous sommes heurtés à plusieurs questionnements dans la modélisation, notamment en ce qui concerne les sous-parties des plantes et les transformations d'ingrédients. Les ingrédients utilisés dans les remèdes ne sont pas toujours des plantes entières. Ils peuvent être des parties de plantes, telles que des graines, des fruits, des racines, des feuilles, etc. Pour modéliser ceci, chaque partie de plante serait un nœud dans un arbre dont la racine est le nom de la plante. Cette manière de modélisation limite le nombre de nœuds et représente efficacement les parties de plantes utilisées dans les remèdes.

Une deuxième question concerne la représentation des ingrédients transformés. En effet, un nombre de remèdes contiennent des ingrédients transformés, modifiant la composition chimique de la plante et altérant potentiellement ses propriétés médicinales. Pour représenter avec précision cette information dans la base de données graphe, une extension du modèle était nécessaire.

L'intégration de ces données dans une base de données orientée graphe correspond à notre besoin de stocker et interroger ces informations complexes. Cette structure de stockage permet non seulement de représenter efficacement les relations entre les entités pharmacologiques, mais elle facilite également des requêtes, rendant ainsi les données plus accessibles.

Finalement, l'application de l'analyse formelle de concepts (FCA) et de l'analyse relationnelle de concepts (RCA) permettent de révéler des patterns et des relations qui étaient auparavant cachées dans les masses de données textuelles[10]. Ces méthodologies permettent de synthétiser

[10] A. Braud, X. Dolques, P. Fechter, N. Lachiche, F. Le Ber, et V. Pitchon, « Analyzing the composition of remedies in ancient pharmacopeias with FCA », in RealDataFCA'2021, ICFCA Workshop, Strasbourg, France, in CEUR Workshop Proc. 3151. 2021, p. 28-35. [En ligne]. Disponible sur: https://ceur-ws.org/Vol-3151/short4.pdf

et de visualiser des connaissances, permettant ainsi de trouver de potentielles nouvelles connaissances sur les pratiques médicinales historiques et leurs contextes d'utilisation. En analysant plusieurs pharmacopées, l'apparition de relations ont le potentiel de contribuer à la biologie et à la pharmacologie en fournissant des bases pour la redécouverte de traitements anciens qui pourraient inspirer de nouvelles recherches pharmaceutiques, notamment la recherche d'alternatives aux antibiotiques.

Table of contents

List of figures

List of tables

Chapter 1: Introduction

1.1 Context

The Abbasid era, which existed between the years 750 and 1258 CE, represents a high point of intellectual and cultural activity in the Islamicate[11] world. This era is distinctive especially for its major contributions to various systems of knowledge, including medical knowledge, which was developed as a result of an integration of medical and theoretical traditions from Greek, Persian, Indian, Syriac and native Arabic knowledge. The Abbasid leadership encouraged a culture of academic inquiry and innovation, with for example the establishment of the House of Wisdom (Bayt al-Hikma) in Baghdad during their period of reign. Indeed, scholars of various traditions came together in the House of Wisdom to translate, research, preserve, and extend the knowledge of medicine [1].

Medical science during the Abbasid period is characterized by systematic efforts towards compiling and improving existing forms of medical literature, especially the production of comprehensive pharmacopeias. These pharmacopeias described a wealth of treatments and cures, while additionally emphasizing a strong code of empirical observation and clinical practice. Established Abbasid physicians, such as Al-Razi (known as Rhazes in the West) and Ibn Sina (known as Avicenna) among the most well-known, made important contributions across a range of domains, from pharmacology to surgical techniques, all of which contributed important principles for West Asian and European medicine for centuries to follow. The pharmacopeias, which are repositories of medicinal knowledge (Figure 1), from this time, are especially important because of the terminology used to catalog and describe the complicated processes of preparation and usage of medical substances. These concoctions showed the tendency of synthesizing remedies from the time-tested traditions and then applying them practically [2].

Nevertheless, the study of these pharmacopeias is not merely of historical interest but has practical modern-day implications, as demonstrated by modern scientific investigations into ancient remedies. Most notably, the re-discovery of a medieval cure, Bald's eyesalve from an Anglo-Saxon pharmacopeia, has shown to have effectiveness against Staphylococcus aureus

[11] Because of the participation of people from other religions and ethnic backgrounds in cultural and academic sphere in this civilization influenced by Arabic culture and Islam, we remain conscious of this throughout the manuscript, and as the lingua franca of the time was the Arabic language, we employ the terms "Arabic pharmacopeias" and "Arabic medicine" with that in mind.

biofilm [3], [4]. Similarly, the work of Tu Youyou, a Nobel Prize-winning researcher who developed the anti-malarial remedy from artemisinin out of ancient Chinese medicine [5], shows there is a role for ancient knowledge to help solve modern health challenges. Indeed, one of the main questions that is of interest entails finding out if potential alternatives to antibiotics can be found from exploring this old knowledge and in less time-consuming ways as opposed to manual analysis, as today's antibiotics are soon to lose their efficacy due to bacterial resistance.

Figure 1 - Kitāb al-bayān fī kashf 'asrār al-ṭibb lil-'iyān, or The Unveiling of the Secrets of Medicine for All, written by Al Ḥamawī, a physician in the 13th century [6] - National University Library of Strasbourg, MS 4187, fol. 10 - Strasbourg, France

1.2 Objectives

To really study the potential insights lying within old Abbasid pharmacopeias, we need to consider strategies dealing with the linguistically difficult, and complicated nature of the texts, as opposed to clear and concise corpora that are the tendency of today's medical knowledge. Thus, our work that aims to develop a strategy for analyzing old Abbasid pharmacopoeias using natural language processing strategies for the benefit of historical and medicinal studies, is multidisciplinary in nature. Its primary goal is to discover valuable knowledge that could contribute to the development of new medications as well as understand the historical aspects of Abbasid physicians. To achieve this, this work involves handling large amounts of unstructured textual data, representing it in a structured manner, and subsequently developing a querying system that allows for the discovery of new knowledge.

Thus, the main question posed by the nature of our work is as follows: How can we automatically convert historical, unstructured textual data in natural language into structured, exploitable data?

Hence, this thesis seeks to implement and adapt a set of approaches to define a pipeline that is applied to the pharmacopeic corpora, extracts useful data, structures it, and then enables querying. Therefore, this work falls within the field of Text Mining and makes use of natural language processing methods to achieve its goals, as Text Mining is the process of deriving meaningful patterns and relationships from large volumes of textual data [7]. Text Mining has been applied extensively in the literature, such as in modern medicine where it is used to uncover insights from modern corpora ([8], [9]) as well as in historical studies, as it facilitates the identification of significant events, influential figures, and historical trends from old documents [10]. However, this is not the case for the study of old pharmacopeias. Thus, this work seeks to explore this area with the goal of valorizing old knowledge in the context of modern research. Figure 2 illustrates the chain of tasks required to complete the pipeline of text mining the old pharmacopeias that we study in this dissertation.

At the heart of this work of analyzing pharmacopeic corpora lies contributions in (1) the processing of named entities to detect them, (2) linking them to their entries in data infrastructures and (3) modeling the available information to (4) perform queries. This indeed makes our main thesis of Text Mining at the intersection of Natural Language Processing, Information Modeling and Data Interrogation, designed to fit the purposes of research in biological studies and historical studies to gain insights into the ingredient combination

strategies of the Abbasid era, on one hand, and to process vernacular terms as to view them in the lens of modern scientific language, on the other hand.

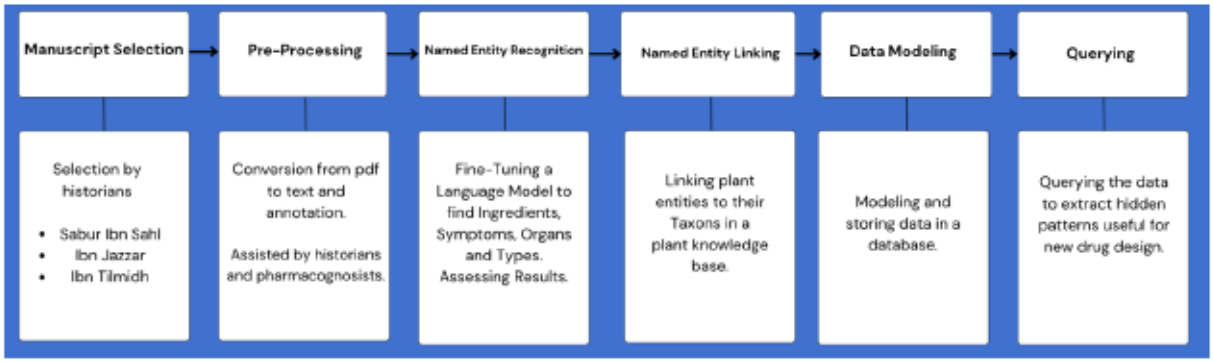


Figure 2 - The chain of tasks in the pipeline of text mining old pharmacopeias.

1.3 Outline

The structure of this dissertation is designed to present the research contributions in a chronological and step-by-step manner, with each chapter building on the previous one to navigate through the various steps of the Text Mining pipeline that was developed. The organization of this dissertation reflects the progression of the research with the following structure:

- Literature review and methodological foundations: This chapter provides an overview of the existing literature on Named Entity Recognition, Named Entity Linking and Disambiguation, and Information Modeling. The review examines past research and methods in these areas and helps guide the selection of the most appropriate approaches for the specific challenges presented by old Abbasid pharmacopoeias. As such, it lays the groundwork for the choices made for the development of the subsequent tasks.
- Contributions in Named Entity Recognition: In this chapter, we present our contributions that focused on the development and assessment of a Named Entity Recognition resource through the fine-tuning of Large Language Models in a manner specifically adapted to the content that is characteristic of pharmacopeias. Multiple models were trained and evaluated, with the DeBERTaV3 [11] model emerging as the most effective for this task, and a study of the generalizability of the recognition task was done in order to understand

how the model behaves across corpora of different writing styles. Additionally, this chapter introduces a new, freely accessible resource that includes the trained model which is available to the community through publication.

- Contributions to Named Entity Linking and Disambiguation: This chapter details the contributions in linking plant ingredients mentioned in the pharmacopoeias to their corresponding taxa. In this chapter, we discuss and critically assess the methodologies employed and the challenges of dealing with vernacular plant names and we present the results of the linkage of these entities to taxonomic infrastructures. We present a novel method of candidate-selection through using geographic references of taxa.
- Contributions in Information Modeling and Querying: This chapter turns to the challenge of structuring the extracted data from the previous task within a graph database, which allows for the detailed representation of relationships between ingredients, organs, remedy forms and symptoms. This chapter explains the process and the difficulties of modeling the data and emphasizes utility of graph-based structures to perform such a task. Furthermore, it shows how combining Formal Concept Analysis and Relational Concept Analysis can be used for the extraction of insights from the structured data and demonstrates how this approach can reveal previously hidden patterns and relationships in the pharmacopeias.

The dissertation concludes with a synthesis of the main contributions and reflects on their significance on the practical aspects of text mining old pharmacopeias and potential future research directions.

Chapter 2: Literature Review

2.1 Named Entity Recognition

2.1.1 Introduction

The task of Named Entity Recognition (NER) holds a central role within the broader field of Text Mining. NER is a task whose objective is to extract useful named entity occurrences from a given text. These elements are lexical units (tokens or a groups of tokens) that refer to things ranging from personal names, organization names and locations to dates, quantities of any kind and specialized vocabulary. The NER task involves two complementary sub-processes: (1) identifying these units in a text, (2) categorizing them contextually according to class types predefined in the task (as shown in the examples in Figure 3). The end-goal and result of the NER task is the annotation of a given text where each recognized entity is tagged with its respective category.

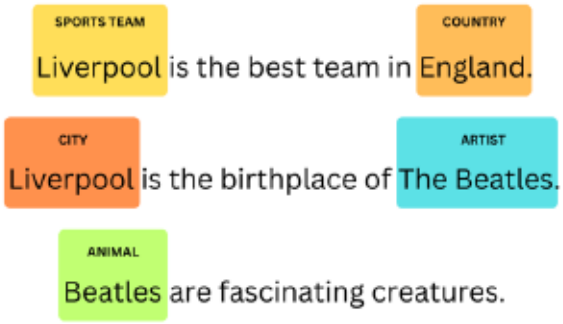


Figure 3 - Example of NER where each entity is tagged based on the context in which they occur.

The applications of this task can include multiple domains, and the focus of this section is on its specialized application to historical documents as well as the medical field. This area has gained increasing scholarly attention due to the unique opportunities it presents [12]. The digitization of ancient manuscripts and historical archives has been an enabler for the application of different techniques within Natural Language Processing, including the task of entity recognition. The initial approaches to NER in this context were predominantly hand-crafted and rule-based, manually adapted to the specificities of the processed corpora. These algorithms were specifically created to accommodate the linguistic and orthographic peculiarities of historical languages [13], [14]. However, these rule-based methods often

proved to be inadequate for several reasons. Firstly, the complexities and variations in ancient texts, including inconsistent spelling and grammatical structures, posed difficulties when it came to the creation of rules. Secondly, these methods were not scalable and required manual intervention for each new corpus or language variety.

The transition from rule-based methods to machine learning algorithms was indeed significant to the field. Traditional machine learning algorithms such as Conditional Random Fields and Support Vector Machines offered a more scalable solution. These algorithms were capable of learning from the data and reduce the need for manual rule creation. However, they still required extensive feature engineering, which was also challenging given the inconsistencies in historical texts [15].

The arrival of deep learning has further revolutionized the Natural Language Processing tasks such as NER. Transformer-based models, such as BERT, have demonstrated superior performance for NER. The ability of these models to take into account contextual information made them well-suited for the complexities of medical texts [16].

This section aims to go more into the details of the methodologies, evaluation metrics, and domain-specific opportunities in NER for our specialized documents. The objective is to provide an understanding of the current state of the art. The section will describe ongoing research and will emphasize the need for domain-specific adaptations [17]. Additionally, this section will study the significance of the "named entity" within the NER task and shed light onto its central role within NER as well as within the processing of natural language.

2.1.2 Understanding Named Entities in NER

What are Named Entities?

Named Entities, in Named Entity Recognition, are the nouns or phrases of the text that hold significant meaning and are of interest for further analysis or processing [18]. We describe them as the focal points around which the linguistic context or semantics of a text revolve. Named entities can be as simple as a single word like "aspirin" or as complex as a multi-word expression like "chronic myeloid leukemia" [19].

Types of Named Entities

Named entities can be categorized into various types based on their semantic roles. The most common types include:

- **Nominal Named Entities**: They are proper nouns, nouns, and phrases such as names of people, organizations, and locations [20]. For example, "Marcus Aurelius" or "The National Museum of Beirut" would be considered as nominal named entities.
- **Temporal Named Entities**: Such named entities that represent time, such as dates, days, and years. For example, "5th century BCE" is a temporal named entity.
- **Numerical Named Entities**: These entities include any numbers or numerical information. For instance, "500 mg" in a medical prescription would be a numerical named entity.

In the context of ancient medical texts, named entities are domain-related and could be specific terms related to medical conditions, treatments, or ingredients. For example, "pastille" or "Scrofula" would be domain-specific entities [21].

Named entities can also be hierarchical or nested within other entities. For example, in the phrase "He was suffering from pain in the stomach" both "stomach" and "pain in the stomach" are named entities, but one is contained within the other (Figure 4). This would be especially challenging in NER and requires specialized models to accurately identify such complex named entities [19].

Figure 4 - Example of an entity nested within another.

Having a semantic understanding of named entity occurrences is necessary in the context of analyzing ancient medical texts for several reasons. First of all, these texts often use archaic or domain-specific terminology that may not be immediately recognizable as named entities. Secondly, the entities in these texts are often the way to understanding the medical practices, beliefs, and knowledge of ancient civilizations, which implies the necessity of their correct identification [3], [4].

Additionally, named entities in NER are not isolated; they often have semantic roles and relationships with other entities or concepts in the text through co-occurrence. For example, in a medical text that says, "green tea alleviates headaches", "green tea" is an ingredient entity,

and "headaches" is a symptom entity. This example indicates that ingredients may be present in the context of sentences that contains mentions of symptoms or diseases. These co-occurrences of named entities can be varied and encompass different types of interactions between entities. In the context of medical texts, entities such as ingredients, symptoms, form, or organs often interact in specific ways that reflect medical knowledge and practices. For instance, a text might describe how a particular herb is used to treat multiple symptoms, or how a combination of ingredients is recommended for a specific condition. As any word can be categorized and assigned to a tag in any given sentence (objects, adjectives, verbs, units of measurement), which would unnecessarily complicate automated NER, it is necessary to identify and carefully define the most important categories for performing NER to yield optimal results.

Another feature of named entities is that they often have multiple meanings or can refer to different things based on the context. For example, "Mercury" could refer to the planet, the chemical element, a singer, or the Roman deity. Semantic ambiguity is a challenge in NER, especially in historical texts where the meaning of words may have evolved over time [20].

2.1.3 Evaluation metrics

Evaluation metrics are essential for assessing the performance of Named Entity Recognition models. They provide quantitative measures that help compare different models and approaches to examine if the chosen model performs well on the task [22]. Common evaluation metrics for NER consist of values of precision, recall, and F1-score where each one offers insights into different aspects of the model's performance.

Precision, for a given category, is the ratio of correctly identified named entities of that category to the total entities of that category identified by the model. It is calculated using the formula:

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

True Positives are the entities that the model correctly identifies as belonging to the category in question. For example, if the model correctly identifies the word "saffron" as an ingredient, it counts as a true positive for the "Ingredient" category. False Positives are the entities that the model incorrectly identifies as belonging to the category. For example, if the model incorrectly identifies the word "saffron" (which is an ingredient) as an organ, it counts as a false positive

for the "Organ" category. High precision indicates that the model makes few false-positive errors, meaning it has a high degree of accuracy when it identifies an entity.

Recall, for a given category, is the ratio of correctly identified named entities of a category to the total actual entities of that category in the data. It is calculated using the formula:

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

False Negatives are the entities that belong to the category in question but are not identified by the model. For example, if the model fails to identify the word "saffron" as an ingredient, it counts as a false negative for the "Ingredient" category. High recall indicates that the model successfully identifies most of the named entities within a category, which means it has a high degree of completeness.

F1-Score is the harmonic mean of precision and recall that provides a single metric that balances both aspects. It is calculated using the formula:

$$\text{F1Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The F1-score is useful when evaluating a model's performance because it provides a single metric that takes into account both accuracy and completeness.

2.1.4 Rule-Based Approaches for NER

Rule-based approaches in Named Entity Recognition (NER) have been around since the foundation of the field and associated with for specialized corpora which include medical or historical texts. These approaches rely on hand-crafted rules, regular expressions, and domain-specific lexicons to identify entities in text. Gorinski et al. [23] demonstrated that a hand-crafted rule-based system was the most accurate way, in their context, to automatically label Electronic Health Records (EHR) for brain imaging reports related to stroke.

In their work, they used EdIE-R (Edinburgh Information Extraction for Radiology reports), which is a rule-based system designed for extracting information from radiology reports [24]. The system processes raw input text through a pipeline that includes sectioning, tokenization, sentence splitting, and linguistic annotation such as part-of-speech (POS) tagging and shallow syntactic analysis. Figure 5 illustrates the EdIE-R pipeline.

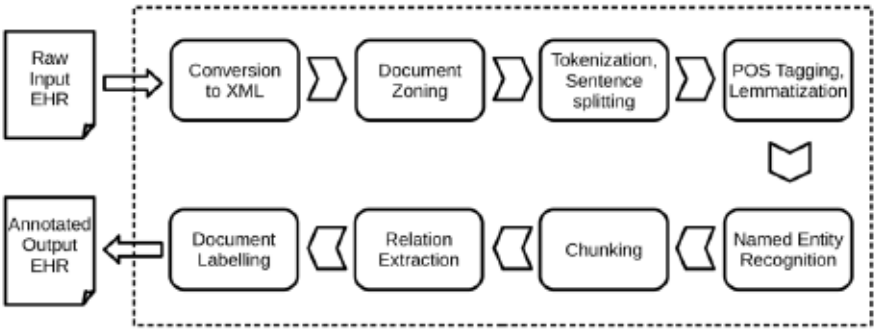


Figure 5 - Overview of the EdIE-R pipeline as adressed by [23].

In the NER step, the system uses hand-crafted rules and lexicons of a specialized vocabulary developed in collaboration with radiology experts. These rules, along with the information from previous steps (tokenization and Part-Of-Speech tagging), rendered possible the identification by EdIE-R of specific target entities. Originally, EdIE-R was able to recognize named entities in brain imaging reports. However, it was concluded that adapting EdIE-R to different datasets, such as radiology reports for other body parts, diseases, or other types of text records, could be costly and time-consuming due to its reliance on hand-crafted rules. Nonetheless, this work showed the suitability of a rule-based approach in highly specialized contexts, in this case modern medicine.

Indeed, rule-based techniques are highly domain-dependent. We see that a main component for building a rule-based system is the use of manually curated domain-specific lexicons. For instance, Kogkitsidou and Gambette [25] focused on old French texts and evaluated the impact of manual and automatic normalization before applying rule-based NER methods. They found that manual normalization led to better results for all methods. Their work suggests that the quality of the lexicon and the normalization process can influence the performance of rule-based NER systems in historical texts.

A primary technique in rule-based NER involves the use of linguistic patterns and regular expressions. These patterns can be simple, such as capitalization rules, or complex, involving part-of-speech tagging and syntactic parsing. As such, rule-based systems rely on a set of predefined linguistic patterns that are manually created to extract the structure and form of entities within a text. Küçük and Yazıcı [26] present a rule-based NER system for Turkish texts, which employs a set of lexical resources and pattern bases for the extraction of named entities. Their system is designed to extract names of people, locations, organizations, as well as time/date and money/percentage expressions from various genres, including news texts, child stories, and historical texts. The domain-specific nature of rule-based NER systems often

means that they are adapted to the characteristics of the target text genre, language and linguistic style. As such, these systems may experience performance degradation when applied to text types that differ from the target domain.

A challenge in applying rule-based NER to historical medical texts is the correction of errors introduced by Optical Character Recognition (OCR) which is a necessary step that precedes the NER task and lies within the pre-processing step. Indeed, if errors occur during this step, some entities might not match the requirements to be recognized as a named entity by the rule-based system. Thompson et al. [27] developed a customized OCR correction strategy for historical medical documents, combining rule-based correction of regular errors with a medically-tuned spell-checking strategy. Their method improved the word-level accuracy of poor-quality documents by up to 16% which notes leads to the improvement of the following NER processes.

Despite its shortcomings, rule-based systems may also prove to be successful for historical texts. McDonough et al. [28] evaluated rule-based NER systems for processing historical corpora which focuses on the geography entries of the eighteenth-century encyclopedia. They found that annotating nested entities (entity within another) and extended place information (with descriptive expansions) improved the performance of their early modern geographic text analysis. This insight is relevant for ancient medical texts, where geographical references to the origins of plants or minerals used in treatments could be important.

Additionally, Kettunen et al. [29] reported on the first large-scale trials and evaluation of NER with data from a digitized Finnish historical newspaper collection. They used a rule-based tagger of Finnish, FiNER, and showed that despite the OCR errors and the noisy data, the rule-based approach was suitable for identifying named entities. This work demonstrates the scalability of rule-based NER approaches, even when applied to large and noisy historical datasets in this case. This may be relevant to inspire a rule-based strategy to perform the NER task on the original Arabic language that is characteristic of the Abbasid pharmacopeias, since Machine Learning approaches, that will be presented in the following section, have their shortcomings in under-resourced languages such as medieval Arabic compared to the well-established state-of-the-art of the English language that is the language of our corpora.

Challenges and Limitations

The literature has shown us that rule-based approaches are highly accurate when well-crafted. Nonetheless, they are also labor-intensive and may not generalize well across different sub-

domains or writing styles. Wen et al. [30] noted that medical NER methods that include rule-based approaches do not make full use of unlabeled medical texts, which can be a limitation especially when dealing with ancient or less-studied languages. They proposed a medical NER approach based on pre-trained language models and a domain dictionary, achieving high F1-scores on unlabeled medical texts. Similarly, Soomro et al. [31] proposed a rule-based approach for biomedical named entity recognition that focuses on disease names. Their work suggests that rule-based classifiers, when combined with statistical machine learning, can yield high precision and recall.

Subsequently, such works demonstrate the potential for integrating rule-based approaches with other techniques to address their limitations, as rule-based approaches in NER have shown high adaptability in specialized domains. These methods offer high precision but require a lot manual effort for rule generation and maintenance, as well as domain expertise.

Despite that rule-based systems prove to be a viable option to examine for NER in old medical texts, we believe that the Abbasid pharmacopeias we work on, even though they are translated to English, contain a highly variable vocabulary which would make it difficult to create exhaustive hand-crafted lexicons that encompass everything related to ingredients, symptoms, organs and forms of remedies. We believe that other techniques, such as the ones within the framework of Machine Learning, might be more suitable for our NER needs.

2.1.5 Machine Learning Approaches for NER

A second strategy for Named Entity Recognition involves Machine Learning. The existing methods are distinguished based on the type of data used for training the different models. Supervised Learning (SL) makes use of only labeled data to develop a model, while Semi-supervised Learning (SSL) combines both labeled and unlabeled data for training. Unsupervised Learning (UL) operates without any labeled data.

Supervised Learning specifically uses text corpora that are pre-annotated or marked for training models. This labeled data, often prepared by human experts, is referred to as training data or the gold standard. Currently, SL is the predominant method used in NER and is able to perform well given sufficient high-quality training data is available. In the following, we will show an overview of earlier statistical methods within SL and then study the details of the more recent Deep Learning methods that became the state-of-the-art of Machine Learning for the NER task.

SL employs models, including:

- Hidden Markov Models (HMMs): Such models are statistical models that assume that the system being modeled is a Markov process with unknown parameters which is a sequence of possible events where the probability of each event depends on the state attained in the previous event only. The HMM is characterized by a series of hidden states, each of which has a probability distribution over the possible observable outputs. This kind of model transitions between states with probabilities, and each state produces an observable output with a specific probability. When it comes to the NER task and making use of the sequential nature of natural language, HMMs help predict the entity category of a word based on the entity categories of adjacent words through two main components: the transition probabilities, which represent the likelihood of moving from one entity category to another, as well as the emission probabilities which represent the likelihood of a particular word of being generated by a specific entity category [32], [33].
- Decision Tree Models: This kind of models uses a tree-like graph of decisions and their possible consequences. Each node in the tree represents a feature of the input data, and the branches represent decision rules that split the data based on these features. The process continues until the data is divided into subsets that are as homogeneous as possible regarding the target variable. The final nodes (leaves) represent the final classification or decision outcome. As such, each path from the root node to a leaf represents a series of decisions that classify the input data into a specific category [34], [35].
- Maximum Entropy Models: This kind of models are based on the principle of maximum entropy. This principle states that without additional information, the probability distribution that best represents the current state of knowledge is the one with the highest entropy. More practically, Maximum Entropy models estimate the probabilities of outcomes given a set of constraints derived from observed data. These constraints are often in the form of features that take into account various properties of the text. Maximum Entropy models are suitable in situations where multiple properties or features influence the understanding of a text. These models can make good predictions even when dealing with complex and high-dimensional data because they consider all possible outcomes and weight them according to the features. For the NER task, these models allow for accurate classification of words in the text by using contextual

information and feature functions to predict the likelihood of different entity categories [36].

- Support Vector Machines (SVMs): Such models function by finding the optimal hyperplane that separates different classes in the feature space. This hyperplane maximizes the margin between the classes. This would help in making the classification robust in relation to variations in the data. Once again, when it comes to NER, SVMs are suitable when there are text features such as word embeddings (which are the representation of words in a vector space), part-of-speech tags, and surrounding context in the text [37].
- Conditional Random Fields (CRF): CRFs are a class of statistical modeling methods often used for tasks involving structured prediction. Unlike models that make independent predictions, CRFs predict a sequence of outputs that depend on a sequence of inputs. In NER and thanks to the sequential nature of human language, CRFs are useful because they consider the context in which a word appears, as well as the dependencies between the labels of adjacent words [38], [39], [40].

The training process is deemed supervised because it involves experts who label the data, guiding the program to make correct distinctions. However, a major limitation of SL is the extensive requirement for high-quality annotated data, which is often scarce and costly to produce. This limitation has prompted the exploration of alternatives such as Unsupervised Learning and Semi-supervised Learning. USL does not depend on annotated data; instead, it forms clusters based on similar contexts, utilizes lexical resources, applies lexical patterns, and analyzes statistics from extensive unannotated corpora. Often, SL systems improve their effectiveness by incorporating USL methods, which utilize unsupervised word representations learned from large volumes of unlabeled data to refine the accuracy of supervised NER models trained with a limited amount of annotated data.

Ultimately, the arrival of Deep Learning has revolutionized the field of NER through addressing the issues of data scarcity and bringing about advancements in model performance and the ability to handle complex language tasks.

The Advent of Deep Learning

Deep Learning is a subset of machine learning that involves the use of neural networks with multiple layers, known as deep neural networks, to model complex patterns and representations

in data. These layers of neurons work in a hierarchical fashion, with each layer progressively extracting higher-level features from the raw input.

Here, we discuss the transformative impact of Deep Learning on NER through exploring innovations such as Recurrent Neural Networks, Long Short-Term Memory networks, Convolutional Neural Networks, Transformers, and disentangled attention mechanisms. These advanced neural architectures have improved NER models' ability to better deal with contextual information, manage long-range dependencies, and classify entities with higher accuracy. The literature suggests that Deep Learning techniques have been able to overcome many of the limitations posed by traditional machine learning approaches and aided in providing more robust and scalable solutions for NER in diverse and noisy text corpora.

One of the main innovations within Deep Learning are Recurrent Neural Networks (RNNs). RNNs are a class of neural networks that are thought for sequence modeling tasks. Regarding the NER task, RNNs can detect the sequential dependencies between words, which is important for accurate recognition of named entities. Long Short-Term Memory (LSTM), a variant of RNN, has been shown to be suitable in clinical entity recognition. Liu et al. [41] employed LSTM for clinical entity recognition and protected health information recognition, achieving F1-scores of 94.37%. A LSTM model consists of three layers: an input layer that generates word representations, an LSTM layer that takes into account the context, and an inference layer that makes tagging. Figure 6 illustrates the neural network architecture designed for the NER task of the cited work. It involves multiple layers to process raw sentences and produce a sequence of labels indicating the entity categories of words in the sentence.

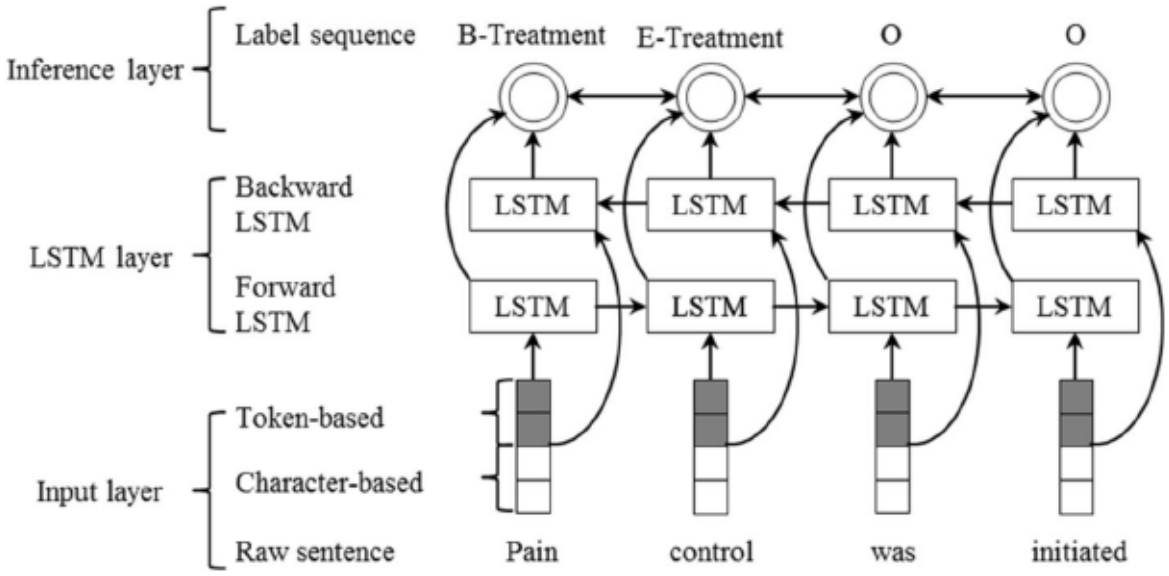


Figure 6 - Overview of the architecture of the LSTM model presented by [41].

In this architecture the first layer is known as the Input Layer. This is how its components are presented:

- Raw Sentence: The input begins with a raw sentence. This example, it includes the words "Pain", "control", "was", and "initiated".
- Character-Based Embeddings: For this case, each word is broken down into its constituent characters. These characters are then processed to produce character-based embeddings, which extracts the morphological structure of each word.
- Token-Based Embeddings: Each word is directly converted into a token-based embedding. This is done to represent the word in a high-dimensional vector space based on its semantic meaning.

The components of the second layer, the LSTM, are defined as:

- Forward LSTM: The token-based and character-based embeddings for each word are provided to a forward LSTM network. This network processes the sequence of words from left to right. The LSTM detects the contextual information from the preceding words.
- Backward LSTM: At a simultaneous time, the embeddings are fed into a backward LSTM network that is to process the sequence of words from right to left. This is done in order to extract the contextual information from the succeeding words.

The outputs from both the forward and backward LSTMs are concatenated to form a joint contextual embedding for each word that integrates information from both directions. This makes the model bi-directional in such case.

The concatenated outputs from the LSTM layer are provided to the inference layer, where the final classification takes place. Its main component is the label sequence. In this case, each word is assigned a label indicating its entity category based on the context provided by the LSTM embeddings. In this example:

- "Pain" is labeled as "B-Treatment" (Beginning of a treatment entity).
- "control" is labeled as "E-Treatment" (End of a treatment entity).
- "was" and "initiated" are labeled as "O" (Outside any entity).

Further, a different innovation within Deep Learning is the development and application of Convolutional Neural Networks (CNNs). These were originally designed for image processing, but they have proven to be effective in tasks such as NER in the literature. The strength of CNNs lies in their ability to detect local patterns and hierarchical features through convolutional filters, which can be adapted to analyze text data by treating it similarly to spatial data.

Regarding the NER task, CNNs can be used to detect and classify named entities within a text by extracting local features such as character n-grams or word n-grams. The work of Dong et al. [42] presents a CNN-based multiclass classification method for mining named entities from electronic medical records. The approach described in this work indicated a pre-processing of the text to generate word embeddings that represent words as dense vectors in a high-dimensional space. These embeddings were then provided to a CNN model to classify each word or phrase as a specific type of named entity. Figure 7 illustrates a typical architecture of a Convolutional Neural Network (CNN) that [42] based their work on.

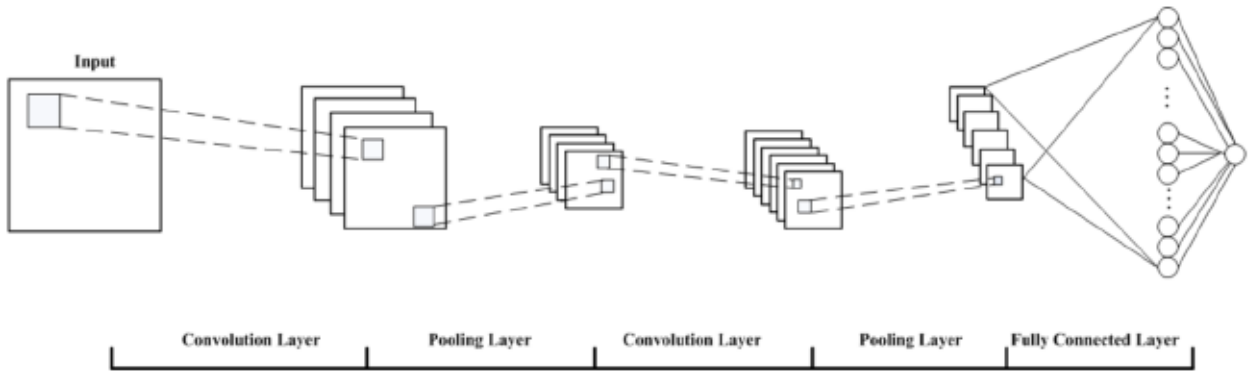


Figure 7 - Overview of the CNN architecture as presented by [42].

In this architecture the first layer is known as the Input Layer. This layer represents the input image, which is usually a matrix of pixel values. The next layer, The Convolution layer, applies convolutional filters (kernels) to the input image. These filters slide over the image to produce feature maps. The output of this layer is a set of feature maps that highlight different features in the input image. Afterwards, the Pooling layer reduces the spatial dimensions (height and width) of the feature maps while keeping the most important information.

This architecture typically alternates between convolution and pooling layers multiple times. Each convolution layer extracts more complex features from the previous layer's output, while each pooling layer reduces the spatial dimensions further.

After several convolution and pooling layers comes the Fully Connected Layer, the high-level feature maps are flattened into a one-dimensional vector and fed into one or more fully connected layers. This is done to combine the extracted features and make predictions based on them. This layer is typically used for tasks like classification, where the output is a probability distribution over different classes which makes it also suitable for the NER task.

One of the more recent innovations in this context is the arrival Transformer models. Unlike RNNs and CNNs, Transformers do not rely on the sequential processing of data, but they make them highly parallelizable. The transformer architecture was introduced by Vaswani et al. [43] in 2017 and is based on self-attention mechanisms that weigh the importance of different parts of the input data. Figure 8 illustrates the architecture of the Transformer model as presented by [43].

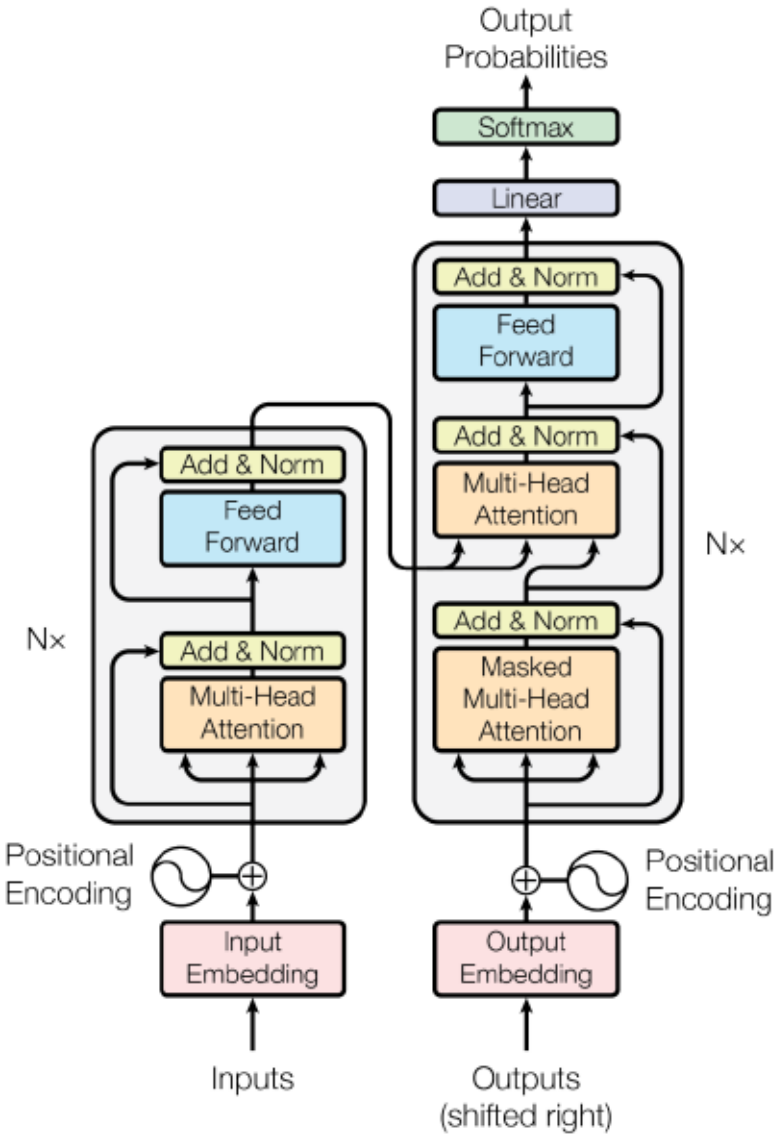


Figure 8 - The architecture of the Trasnformer model as first introduced by [43].

The Transformer architecture is composed of an encoder and a decoder. Both of them are stacks of identical layers. The encoder processes the input sequence and generates an encoded representation. Initially, raw input tokens are converted into dense vector embeddings that represent their meanings in a high-dimensional space. Considering that the Transformer model does not have a built-in sense of order of the sequence, the positional encodings are added to the input embeddings to inform of the position of each token in the sequence.

Afterwards, the input embeddings with positional encoding are passed through a stack of N identical layers. Each layer here consists of two main sub-layers which are Multi-Head Attention and Feed Forward. On one hand, the Multi-Head Attention layer allows the model to focus on different parts of the input sequence simultaneously by applying multiple attention

mechanisms in parallel. Each attention head computes a weighted sum of the input values, where the weights are determined by the similarity between queries and keys. On the other hand, the Feed Forward layer applies a fully connected feed-forward network to each position in the sequence independently and identically. Additionally, residual connections (Add) and layer normalization (Norm) are applied after each sub-layer to facilitate training and stabilize the model.

Afterwards, the decoder generates the output sequence one token at a time, based on the encoded representation from the encoder and previously generated tokens. The target tokens are also converted into dense vector embeddings and added to positional encodings. The first sub-layer in each decoder layer is masked Multi-Head Attention. The masking ensures that the prediction for a particular position depends only on the known outputs up to that position, not on future positions. The second sub-layer performs Multi-Head Attention over the output of the encoder stack, allowing the decoder to attend to all positions in the input sequence. Similarly to the encoder, the decoder layers also include Feed Forward networks and Add & Norm operations. The final output of the decoder is passed through a linear layer followed by a softmax layer to produce a probability distribution over the target vocabulary for each position in the output sequence.

For the NER task, the Transformer model can be used to classify each token in the input sequence into predefined entity categories. The encoder receives the contextual information of the entire sequence, while the decoder generates labels for each token based on this context. The self-attention mechanism in the encoder allows the model to weigh the importance of each token relative to others by detecting dependencies and relationships within the sequence.

In the context of medical texts, transformers have shown success. For example, Schmidt et al. [44] applied transformer-based models for classification and question-answering tasks in clinical trial texts, achieving F1-scores reaching 94%. Although their work focused on modern medical texts, the flexibility and high performance of transformers suggest their applicability to ancient medical texts, especially given the ability of such models to handle ambiguity and insufficient training data.

Indeed, building on the Transformer architecture, BERT (Bidirectional Encoder Representations from Transformers) has been influential in the literature. BERT pre-trains deep bidirectional representations by conditioning on both left and right context in all layers.

Devlin et al. [45] introduced BERT and demonstrated its effectiveness in various NLP tasks, including NER. It is notable to observe that BERT and its variants have not been extensively applied to ancient medical texts in the literature. Nonetheless, their robustness and flexibility make them strong candidates for such specialized domains and suitable for our needs. More recently, disentangled attention mechanisms innovated on the Transformer architecture and aimed to separate different types of syntactic and semantic information in the attention scores. DeBERTa is a model that incorporates disentangled attention mechanisms [11], where the attention mechanism is modified to disentangle content and positional information. This means that separate attention mechanisms are used to process the semantic content of the tokens and their positional encodings. This allows the model to more accurately detect the relationships between tokens based on their content and their positions in the sequence. Indeed, this is shown in the work of Martin et al [46] where when compared to other Transformer models, the DeBERTa-XLarge model ended up having the best overall results (F1-score = 87.2%) and had the highest recall score of 87.6%.

Limitations and Future Directions

Machine Learning techniques have shown promise in NER for medical texts, but they also have limitations. One of the primary challenges in this context is the need for large, high quality annotated datasets for training [47], which are often not available for specialized domains like old pharmacopeias. Furthermore, deep learning models like RNNs and LSTMs are computationally expensive, requiring specialized hardware for training and inference [41].

Another limitation is the interpretability of these models. Indeed, Deep Learning models are often criticized for being "black boxes" because it is difficult to understand the precise details of their decision-making processes [48]. This is a concern in medical applications where interpretability is important for trust and reliability.

Future directions in this area could involve the integration of domain-specific knowledge into machine learning models to improve their performance and interpretability. For example, medical ontologies could be incorporated into the training process to provide contextual information that can help the model better understand the semantics of the text.

In addition to the posed future directions, it is notable to add that the Generative Pre-Trained Transformer (GPT) has been a big advancement in Natural Language Processing, including the NER task. Developed by OpenAI, GPT is trained unidirectionally which makes it suited for generative tasks. It has also been fine-tuned for various classification tasks, including NER

[49], [50]. While GPT has not been extensively applied to old pharmacopeias in the literature, its architecture allows for large-scale unsupervised pre-training, which could be beneficial for domains with limited annotated data.

It is clear that the progression from rule-based methods to advanced deep learning models has improved the capabilities of NER systems. The continuous development of novel architectures and techniques, alongside rigorous evaluation, will further drive improvements in the accuracy, scalability, and applicability of NER across various domains. As the field advances, we believe that it is important to make use of the ever so improving NER task towards the goal of Text Mining old pharmacopeias for insight extraction, as part of a general Text Mining pipeline.

2.2 Named Entity Disambiguation and Linking

2.2.1 Introduction

Following the previous task of NER, and in the context of Text Mining, Named Entity Disambiguation (NED) and Named Entity Linking (NEL) are necessary processes that facilitate the structuring of large volumes of text through providing additional information regarding extracted named entities. Indeed, these tasks are deemed relevant in the aim of extracting meaningful information from textual data sources.

NED involves identifying and resolving ambiguities associated with named entities mentioned in a text. A named entity is defined as discussed in the previous section on NER (2.1.2); the challenge in NED lies in correctly identifying which semantic meaning a particular name entity refers to, especially when multiple named entities share the same or similar names.

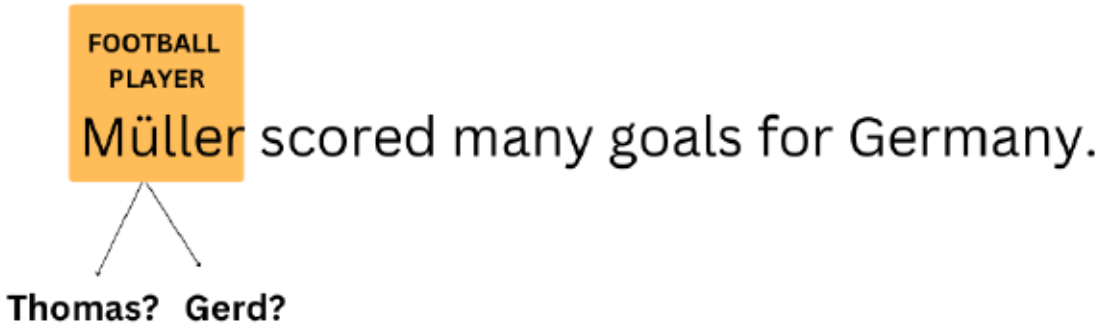


Figure 9 - Example of semantic ambiguity in the context of a recognized named entity

For instance, let us consider a named entity tagged as "football player" during the NER task as in Figure 9. This named entity is "Müller" in the sentence "Müller scored many goals for Germany". We know from the NER task that "Müller" is a football player, however, as this is a common last name in the German culture, we do not know which real-world person it refers to; we describe this as semantic ambiguity. Two of the most likely candidates would be Thomas Müller and Gerd Müller, as these have had the most famous careers in a German football context. The *goal* of the NED task in this situation would be to provide the most likely candidate that the named entity refers to. This would make use of contextual aspects that can lead to a better ranking of the candidates. For instance, if the sentence were "Müller scored many goals for Germany in 2010", the temporal information would provide a cue towards asserting the likelihood of Thomas Müller (who played in 2010) being the named entity in

question, while "Müller scored many goals for Germany in 1974" would imply Gerd Müller. This task is commonly investigated in the literature and makes use semantic knowledge resources, such as encyclopedic knowledge [51] and Wikipedia [52], [53].

NEL extends the disambiguation to link the identified entity to a unique identifier in a knowledge base or database. This process involves mapping the disambiguated entity to an existing entity within a structured set of data which allows for the integration of textual data with additional information for that data. However it seems that the literature seems to consider NEL and NED to be the same task, while others seem to separate them and yet still view them as part of the same goal [7], [14], [54], [55], [56]. In other words, it seems that resolving ambiguity is an integral part of the aim of mapping a named entity to its standardized form in a knowledge base. Indeed, Tedeschi et al. [56] said "Entity Linking (EL), also known as Named Entity Disambiguation (NED), is the task of associating an ambiguous textual mention with a named entity in a knowledge base." which implies that what we are observing as two sub-tasks are the same task.

On the other hand, as posed by [57], the NED task can be seen as a subtask of the NEL task because NED is concerned with disambiguating a textual Named Entity mention, where the correct Named Entity is known to be one of the knowledge base entries, while the NEL task also handles the cases where there is no entry for the entity in the reference knowledge base. In other words, entity linking does not only require the disambiguation to be performed prior to the linking to the target entry of the knowledge base, but it also goes a step further and manages the detection and appropriate handling of entities that are absent from the knowledge base.

The literature seems not to be unanimous in the definition of the relationship between NED and NEL but it is clear that they are steps that serve the same purpose of mapping a recognized named entity to its entry in a standardized data infrastructure. As such they are important within the goal of transforming unstructured text into structured, queryable data. In our work, we consider that NEL implies the task of disambiguation and proceeds further to use that information to actively map the named entity to its entry, or lack thereof, through information modeling as presented in Chapter 5: Figure 10 from [7] demonstrates the entire pipeline from text pre-processing, through NER and NED, to NEL and the construction of a knowledge graph. It shows how raw text can be transformed into structured, linked data that reveals the relationships between entities and enriches the information extracted from the text with knowledge base references, through passing from the NER task, to the disambiguation of found

named entities and their linkage to standardized data infrastructures in order to model the resulting information in a Knowledge Graph that is queryable.

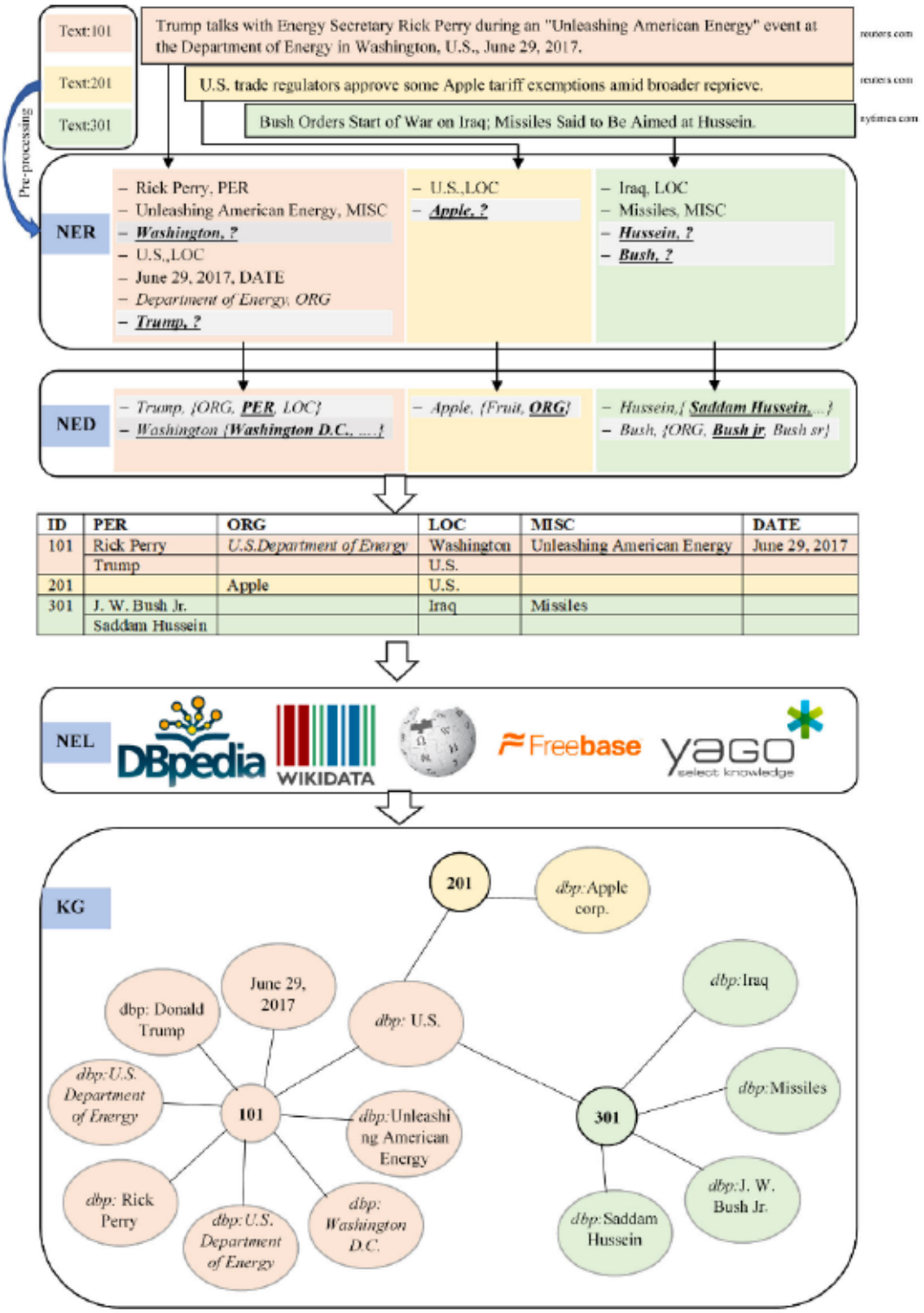


Figure 10 - A pipeline of Named Entity processing through its different steps, from [7].

NED and NEL improve the accuracy of content analysis through identifying and linking named entities, and they facilitate the construction of data models because the information is provided

beyond the simple mention of a named entity. This is relevant in fields such as the study of old pharmacopeias because the extraction of appropriately referenced information from old texts is important for understanding historical contexts and medical practices. Indeed, the primary challenge in NED and NEL is the ambiguity and variability of natural language. Texts often contain incomplete, implicit, or context-dependent references to named entities. The problem is further present in short, noisy texts, such as tweets or headlines, where limited contextual information is available [58].

The evolution of NED and NEL can be seen through the lens of various stages that have marked the advancements in methodologies and technologies of the tasks at hand and aligns with the broader evolution of Natural Language Processing and Computational Linguistics as a whole.

At first, much like the NER task, the initial development of NED and NEL was dominated by rule-based approaches. These methods relied heavily on handcrafted rules and heuristics to identify and disambiguate named entities. The focus was primarily on extracting entities based on syntactic patterns, dictionary lookups and useful metadata. Even though this was effective to a certain extent, these approaches were limited by their inability to adapt to the variability and complexity of natural language. However they remain efficient for limited domains and relevant within the state of the art [58], [59], [60]. The integration with external knowledge bases presented a striking advancement in the later history of NED and NEL. This development allowed for the enrichment of disambiguation processes with semantic information from structured sources like Wikidata or domain-specific data infrastructures. This integration facilitated a more context-aware approach to entity linking, improving the accuracy and reliability of the disambiguation process [61], [62], [63].

Following the rule-based era, the field witnessed a shift towards statistical methods. These approaches utilized probabilistic models to infer the likelihood of an entity being a particular real-world object. Techniques such as Hidden Markov Models and Conditional Random Fields became popular. These methods offered more flexibility and adaptability compared to rule-based systems, as they could learn from annotated corpora and adjust to different linguistic contexts [64], [65], [66]; indeed, there is a striking parallel between the evolution of NED/NEL and the NER task as described in 2.1 because these tasks seem to frequently complement each other in the literature in the same way that they do in our work.

Thus, the most recent phase in the evolution of NED and NEL is also characterized by the adoption of deep learning methods. The introduction of neural network-based models such as

those with Transformer architectures like BERT [45], marked a paradigm shift in this area. As these models are capable of dealing deep semantic representations in the form of vector spaces, this was able to improve the performance of NED and NEL systems. They brought about an end-to-end approach through integrating entity recognition and disambiguation in a single, unified process [67], [68], [69].

In order to address the unique needs presented by specific domains, such as historical texts and medical documents, a detailed overview of the current state of the art in NED and NEL is needed. We examine the latest methodologies and developments in NEL and NED and we aim to gain a clearer knowledge of how these processes can being adopted and refined to meet the specific needs of this thesis.

2.2.2 Rule-Based Approach

The rule-based approach is a traditional method applied in entity disambiguation and linking. It operates under the principle of using predefined rules or heuristics, facilitating the identification and association of named entities within a given text.

The rules applied in this approach are often created by domain experts or linguists, informed by an understanding of the specific entities relevant to the target domain. These rules encapsulate specific patterns, syntactic structures, or contextual indicators in order to align with the attributes of specific entities. The formulation of rules is a process that requires an in-depth analysis of the linguistic and contextual attributes of named entity references in each corpus. The range of these rules can be broad, from simple pattern matching rules (e.g., words that end in "ing") to complex linguistic rules that account for syntactic dependencies (e.g., nouns that follow a preposition), semantic relationships (e.g., words that have high semantic similarity to the word "king" in a vector space), and co-occurrence patterns (e.g., words that are within 3 words of "remedy").

The early rule-based systems for NED and NEL were primarily designed to extract entities based on syntactic patterns, dictionary lookups, and metadata. For instance, Cucerzan [53] demonstrated an approach where Wikipedia was used as a resource for disambiguation and made use of its structured data to improve the accuracy of entity identification. Figure 11 illustrates the different processes that were part of their system of disambiguation and linking to Wikipedia, where the red arrows represent the chronological passing of different tasks within the proposed pipeline and the blue arrows represent the strategies employed to perform said tasks.

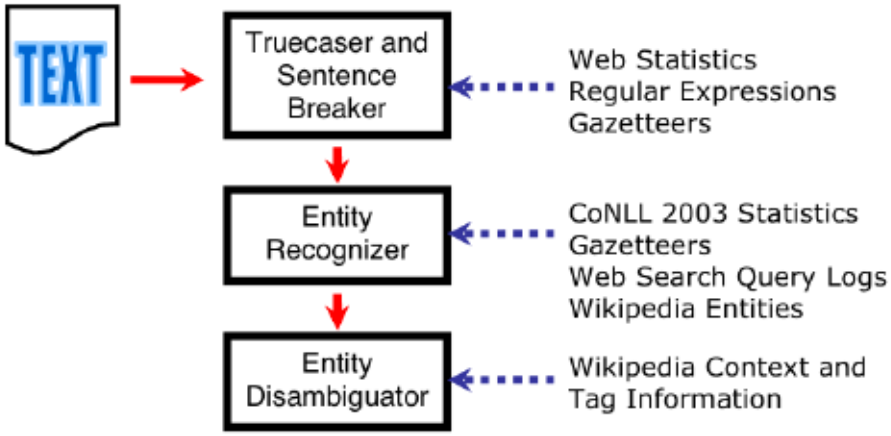


Figure 11 - Overview of the processes employed by the system proposed by [53].

Similarly, the work by Volz, Kleb, and Mueller [70] on ontology-based disambiguation of geographical identifiers shows the importance of domain-specific rules in improving the precision of entity linking. Their approach demonstrated the necessity of adapting rule-based systems to specific domains like geography to address the ambiguities in those fields.

Shen, Wang, and Han [71] provided a overview of entity linking techniques including rule-based methods. They discussed the evolution of entity linking systems and described the transition from rule-based to more advanced approaches. They also show the limitations of early systems in dealing with name variations and entity ambiguity.

The evolution of rule-based approaches has been marked by the integration of more sophisticated techniques to address their limitations. For instance, Srinivasan and Rafiei [58] proposed a location-aware NED framework that integrates spatial signals to resolve ambiguities in short and noisy texts like tweets and news headlines. This approach represents demonstrated how the integration of additional contextual cues such as spatial information, can improve entity disambiguation.

Additionally, the work by Mobasher et al. [59] on combining dictionary- and rule-based approximate entity linking with tuned BioBERT indicated the potential of hybrid approaches. Their system applied a two-stage approach, by firstly using the fine-tuned BioBERT for identification of chemical entities then performing a semantic approximate search in chemical databases for entity linking.

One of the advantages of rule-based systems was their interpretability. Since the rules were explicitly defined, it was easier to understand and debug the system's decisions. However, this strength was also a limitation. Developing these rules required extensive domain knowledge

and manual effort. The rules are often rigid and could not easily adapt to the variability and complexity of natural language, much like we described them in the context of the NER task.

As the field evolved, the limitations of rule-based approaches in handling the complexity and diversity of natural language led to the exploration of more advanced methods. This transition marked the beginning of the use of statistical models and, later, machine learning techniques in NED and NEL. The aim of these approaches was to learn from annotated corpora which allowed for adaptability and accuracy across various text types and domains.

2.2.3 Machine Learning in Named Entity Disambiguation and Linking

The evolution from rule-based to Machine Learning methods in Named Entity Disambiguation and Linking marked a shift in the field of Text Mining. This shift introduced more adaptable systems that made use of probabilistic and algorithmic models to infer the likelihood of an entity being a particular entry in a data infrastructure.

First, statistical methods are characterized by their use of mathematical models to analyze and interpret data and employ algorithms that can learn from the data and make predictions or decisions based on the statistical probabilities.

Hidden Markov Models, as discussed in 2.1.5, are statistical models that excel in modeling sequential data in contexts where the states of the sequence are not directly observable [72]. In Named Entity Disambiguation and Linking, HMMs are employed to model the sequence of words in a text, with each word having a probability of being associated with a particular type of named entity. The strength of HMMs lies in their ability to handle sequences where the actual states (named entities in this case) are 'hidden' and only the observations (words in the text) are visible. HMMs assume that the system being modeled is a Markov process with these unobserved states which makes them suitable for NED and NEL where the goal is to infer the most likely sequence of entities based on observed words [57].

For instance, in the work of Priya [65], HMMs were employed for entity extraction and used n-gram features and parts of speech clustering. The statistical nature of HMM allowed for the modeling of entity sequences in text through providing a probabilistic framework to handle the ambiguity in entity recognition.

Further, Conditional Random Fields (CRFs) that are a type of statistical modeling technique characterized by being discriminative, model the conditional probability of the output (the Named Entities) given the input (the sequence of words in the text). This makes CRFs effective

in NED and NEL tasks where context plays an important role. CRFs focus on modeling the probability of the hidden state (named entity) given the observed state (word in the text), considering both past and future input features. This ability to consider the entire context makes CRFs more adept at capturing the dependencies and relationships between entities in a sequence [73].

A contribution for this area is found in the comparison of features for Part-Of-Speech (POS) tagging in Kannada by Atmakuri et al. [74], where CRFs were used for sequence modeling in POS tagging, a necessary step in NED. CRFs offered flexible approach to model the dependencies between tags in a sequence which makes them suitable for tasks related to dealing with named entities. This paper shows the adaptability of CRFs in handling various linguistic features and described their improvement of the accuracy of POS tagging.

In the disambiguation work of Balaji and Sasikala [75], the focus was on webpage perception through a model of Hierarchical Conditional Random Fields (HCRF). This work addressed the challenge of integrating academic communities for similar interest groups by perceiving their entities using HMM and CRF. The flexibility of HMMs in modeling different aspects of language and the robustness of CRFs in finding hierarchical relationships indicated the usefulness of these statistical methods in NED.

Additionally, other methods rely on non-probabilistic Machine Learning models that have proven useful within the NEL and NED tasks. Support Vector Machines (SVMs) in NED and NEL are noted for their effectiveness in classification tasks. SVMs work by finding the optimal hyperplane that separates different classes in the feature space. When it comes to NED and NEL, SVMs can be trained to distinguish between different types of named entities based on their features like word embeddings and contextual information. This capability makes SVMs suitable for classifying entities into categories with high accuracy and thus aided in candidate selection.

A contribution to the use of SVMs in NED is seen in the work of Alokaili and Menai [76], who studied SVM ensembles for NED. Their study examined the effectiveness of SVM ensembles in accurately classifying entities and described the robustness of SVMs in handling the complexities of NED tasks, by performing best on benchmark corpora AIDA/CONLL-TestB and AQUAINT with F-score respectively reaching 78.5 and 71.5%.

Further, Habib and V. Keulen [77] demonstrated the application of SVMs for NED in social media contexts such as in Twitter (now known as "X" at the time of writing). Their approach

involved using SVM to rank candidate pages for entity disambiguation, which aimed to study the adaptability of SVMs in extracting and disambiguating named entities from tweets.

Decision Trees, another traditional machine learning model, offer a more intuitive approach to entity classification. These models use a tree-like structure where each node represents a feature of the entity, and each branch represents a decision rule. In NED and NEL, Decision Trees can be useful for making decisions based on a series of simple, interpretable rules derived from the training data. This can be advantageous in scenarios where explainability is as important as accuracy. A notable application of Decision Trees in NED is found in the work of Liu, Xu, Lu, and Xu [78]. In their study, they made use of Decision Trees to disambiguate person names in Chinese text, utilizing features extracted from the Chinese encyclopedia Baidu Baike.

Additionally, the research by Wahio, Suzuki, Ting, and Inokuchi [79] included discussions on disambiguation through the use of Decision Trees for various data mining tasks. Their work described the usability of Decision Trees in handling different types of data and their potential in improving the accuracy of entity classification.

However, these methods also faced challenges because of the need for large annotated corpora for training and expensive computational resources. The reliance on such models also meant that the performance of these systems was directly impacted by the quality and quantity of the training data in the same way that we posed in the NER section.

Indeed, drawing a parallel to the evolution of the NER task, the progression from statistical and algorithmic methods to Deep Learning in Named Entity Disambiguation and Linking represents a leap forward in the field. This evolution introduced systems that can adapt to linguistic contexts and improve over time by using the vast capabilities of modern computational models. Deep learning further extends the capabilities of NED and NEL systems by employing neural networks with multiple layers to model complex patterns in data. Deep Learning techniques have brought about improvements in NED and NEL, especially with the arrival of architectures like BERT.

Transformer Models, which are the backbone of architectures like BERT, have revolutionized NED and NEL by enabling the processing of entire sequences of text simultaneously (as opposed to sequentially). This parallel processing capability, combined with attention mechanisms (Figure 8 – section 2.1) that allow the model to focus on relevant parts of the text, has improved the accuracy and efficiency of NED and NEL systems. For instance, unlike

traditional models that process text in a single direction (either left-to-right or right-to-left), BERT is designed to consider the full context of a word by looking at the words that come before and after it. This bidirectional context understanding is important for accurately disambiguating named entities in text, as the meaning of a word can change dramatically based on its surrounding words.

The work of Borchert and Schapranow [80] studied an approach to biomedical NEL in Spanish clinical case reports. The system was developed to extract disease mentions and link them to concepts of SNOMED CT, which is a systematically organized computer-processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. The authors utilized a Transformer-based NER model pre-trained on Spanish biomedical documents to identify disease mentions. For candidate generation, they combined a TF-IDF vectorizer with a cross-lingual SapBERT model. This hybrid approach was further improved by a rule-based reranking step that adjusts candidate lists based on semantic types and other criteria. The system achieved a micro-averaged F1-score of 0.566.

Additionally, Kolitsas et al. [81] introduce a method that jointly addresses mention detection and entity disambiguation within a unified framework. They used a bidirectional LSTM and an attention mechanism and proposed a model that simultaneously discovers and links entities which allows to make use of the mutual dependency between mention detection and entity disambiguation tasks (Figure 12). In this figure, the main essence that we understand from it is that the sentence "The New York Times is an American newspaper" was processed using several layers. Initially, word embeddings and character embeddings for each word are generated. These embeddings are passed through a bidirectional LSTM, capturing context-aware word representations. The resulting embeddings are then fed into a feedforward neural network (FFNN) to generate mention representations. A set of candidate entities is considered for each mention, with each candidate having an embedding. Then, the mention representation is combined with each candidate entity embedding, and a similarity score is computed using another FFNN. The final local score is obtained by integrating these similarity scores. A global disambiguation layer also considers the coherence of the candidate entities within the entire document context and adjusts the final scores accordingly. This hierarchical approach allows the model to disambiguate mentions by considering both local context and global document-level information. This end-to-end approach showed an improved contextual understanding of mentions and improved the accuracy of entity linking.

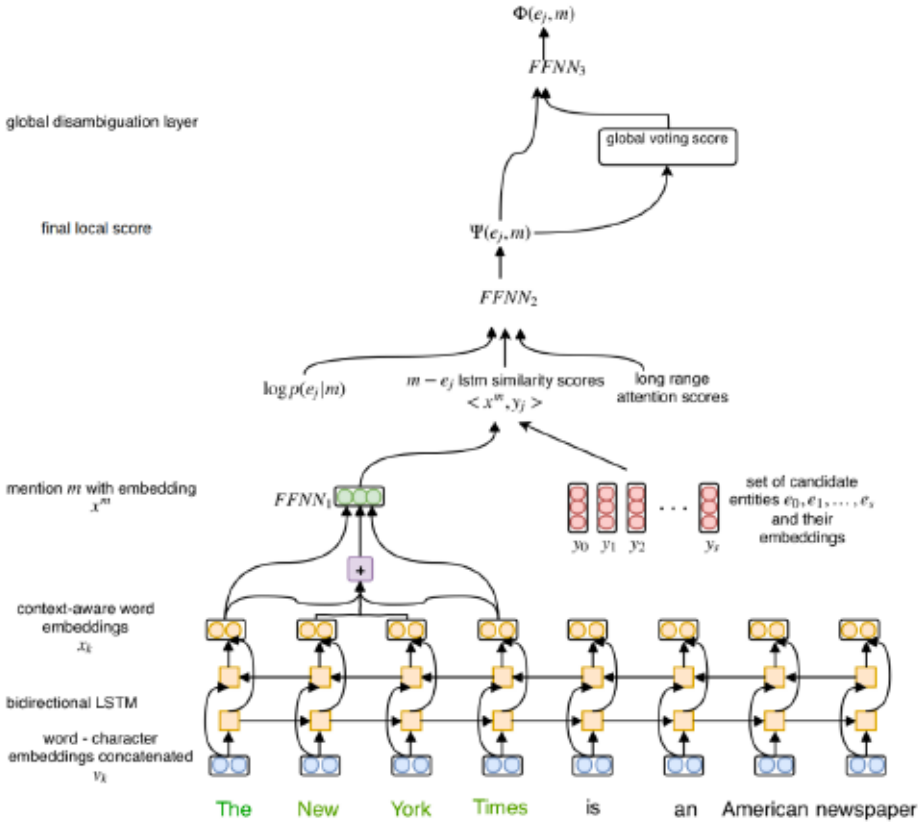


Figure 12 - Global model architecture for the mention "The New York Times" where the final score is used for both of the mention linking and the entity disambiguation decisions, from [81].

Complementing this work, the study by Boros et al. [82] focuses on the challenges posed by historical and multilingual texts. This paper presents a system that employs a transformer-based model for NER and the BiLSTM-based model for Entity Linking presented by Kolitsas et al. [81] (Figure 13), specifically designed to handle the noisy and segmented nature of OCR-digitized historical documents.

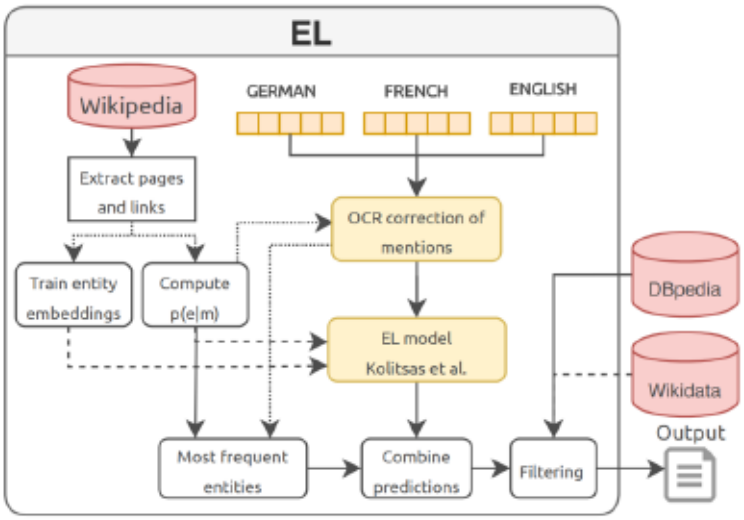


Figure 13 - The Entity Linking (EL) and post-processing model proposed by [82]

Both aforementioned studies show the importance of integrating advanced neural architectures with sophisticated pre-processing techniques to address the complexities of NER and EL in diverse textual environments. While Kolitsas et al. [81] describe a joint learning approach to mention detection and entity disambiguation, Boros et al. [82] extend the application of these techniques to the challenging domain of historical multilingual documents, illustrating the broad applicability and effectiveness of neural entity linking methods. These contributions underscore the potential for further advancements in the field through the continued integration of neural networks and developed data pre-processing strategies.

The integration of machine learning and deep learning in NED and NEL has indeed improved the accuracy and efficiency of these systems and opened new possibilities for handling more complex entity disambiguation tasks. However, much like for its NER counterpart, these techniques create the the need for substantial computational resources and large datasets for training.

2.2.4 The Case of BabelNet

BabelNet is a multilingual semantic network and encyclopedic dictionary that combines lexicographic and encyclopedic knowledge from WordNet and Wikipedia [83]. It is designed to provide an integration of concepts and named entities across multiple languages. Indeed, this would make it a powerful resource for the task at hand.

It integrates data from various sources, including:

- WordNet: A lexical database for the English language.

- Wikipedia: A vast repository of encyclopedic knowledge.
- OmegaWiki: A multilingual dictionary.
- Wikidata: A collaboratively edited knowledge base.

This amalgamation creates an extensive resource that offers both lexical and semantic information about words and entities. BabelNet organizes knowledge into synsets (sets of synonyms) and links them to the corresponding Wikipedia pages and ensures a coverage of both common words and named entities.

One of the distinguishing features of BabelNet is its support for multiple languages. In its latest version 5.3, it encompasses data in over 600 languages, facilitating cross-lingual linking tasks. This is beneficial and relevant for applications dealing with texts in diverse languages because it ensures that entities can be disambiguated regardless of the language.

BabelNet's extensive coverage and multilingual nature make it highly suitable for NEL/NED. It provides word sense disambiguation algorithms that make use of the semantic relationships between entities as to help resolve ambiguities. According to Moro et al. [84], Entity Linking and word sense disambiguation both tackle the issue of lexical ambiguity in language. Despite their similarities, they differ in one aspect: in entity linking, a textual mention is associated with a named entity, which might not always perfectly match the mention itself. In contrast to this, word sense disambiguation involves a direct correspondence between the word form (or its lemma) and the appropriate word sense.

This work demonstrates that the research community has approached NEL and word sense disambiguation as separate tasks, often leading to duplicated efforts and solutions. Contrary to this trend, Moro et al. [84] proposes a move towards efficiently integrating encyclopedic and lexicographic knowledge into structured language resources. Such structured resources naturally suggest a common ground for both sense disambiguation and entity linking tasks. Specifically, this work investigates the hypothesis that lexicographic knowledge used in word sense disambiguation can be beneficial for NEL, and conversely, that encyclopedic information used in NEL can aid in disambiguating nominal mentions.

Babelfy: A Unified Approach to Entity Linking and Disambiguation

Babelfy is a state-of-the-art system built on top of BabelNet, offering a unified approach to word sense disambiguation and NEL. It makes use of the wide semantic network of BabelNet to perform disambiguation and linking tasks efficiently [84], [85].

Babelfy employs a graph-based approach to disambiguation (Figure 14). It represents the input text as a semantic graph where nodes correspond to BabelNet synsets and edges represent semantic relations between them. The algorithm proceeds through the following steps:

1. Candidate Selection: For each ambiguous word or entity in the text, Babelfy identifies potential candidates from BabelNet.
2. Graph Construction: It constructs a semantic graph by linking candidates based on semantic relations in BabelNet.
3. Disambiguation: Babelfy uses a centrality-based heuristic to select the most coherent subgraph to disambiguate the entities and linking them to the appropriate BabelNet synsets.

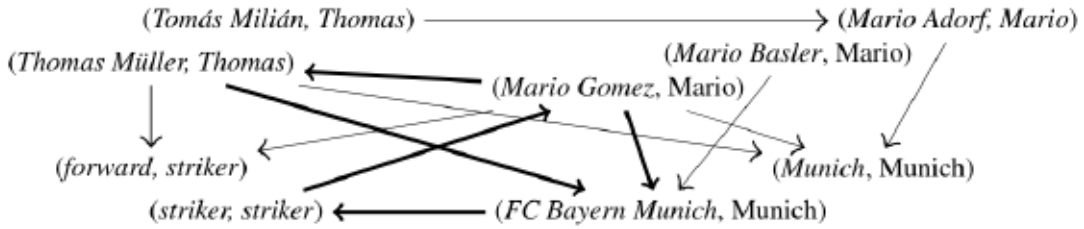


Figure 14 – Example of the semantic interpretation graph built for the sentence "Thomas and Mario are strikers playing in Munich" where the edges connecting the correct meanings are in bold, from [84].

This process ensures that Babelfy can handle both NEL and word sense disambiguation in a unified manner by using the interconnectedness of BabelNet's semantic network. This tool has demonstrated high performance across various NEL benchmarks. Its ability to make use of BabelNet's extensive and multilingual knowledge base allows it to disambiguate entities accurately even in noisy or contextually complex texts. In the context of ancient pharmacopeias, Babelfy can link mentions of plants, symptoms, and ingredients to their dictionary forms and synonyms and data infrastructure instances and aid in the NEL task and thus in the understanding and analysis of these documents.

Despite its strengths, however, Babelfy faces challenges in dealing with extremely noisy text or texts with insufficient context. The reliance on BabelNet also means that any gaps or inaccuracies in the knowledge base can impact Babelfy's performance. The quality of disambiguation can vary depending on the richness of BabelNet's coverage for a particular language.

We believe that Babelfy and BabelNet have set a state-of-the-art standard for NEL/NED tools through integrated solutions that use extensive multilingual data infrastructures, which is why this tool is useful for the context of our work on old pharmacopeias and will be used in our study.

2.3 Information Modeling in Text Analysis

2.3.1 Basics of Data Modeling

Information modeling is an important aspect of information representation and manipulation. It is relevant when dealing with named entities as they need structured representation for efficient processing and analysis. Data modeling provides framework that is needed to organize these entities systematically and ensures that their retrieval and manipulation are. What we mean by Information Modeling is the process of creating a data model for the data to be stored in a database. This model defines the logical structure of the data, including the relationships and constraints that govern how data can be stored and manipulated [86], [87]. In the context of named entities, information modeling helps in structuring the extracted entities in a way that improves their usability and accessibility for further analysis.

Data models can be categorized into three main levels of abstraction which are conceptual, logical, and physical [88]. First of all, Conceptual Data Models are representations that describes the structure of the data, and they are primarily used to define the overall structure and relationships within the data. Secondly, Logical Data Models provide a detailed view of the data structure and specifies the types of data, relationships, and constraints and include detailed attributes, primary and foreign keys, and normalization processes. Finally, Physical Data Models represent the actual implementation of the data structure within a database and include specifications for things such as tables, columns and indexes and they translate the logical design into an instance that can be used to create and maintain the database.

Early data modeling efforts focused on hierarchical and network models, which were limited in their flexibility and scalability. The introduction of the relational model by Codd in the 1970s created a paradigm shift in data modeling by providing a more flexible and intuitive way to represent data using tables and relationships [89].

The relational model's introduction led to the widespread adoption of relational databases, which became the standard for data storage and management. However, as data complexity and volume increased, new models and technologies emerged to address the limitations of the relational approach. More recently, graph databases have provided robust solutions for representing complex relationships and interconnected data [90].

In our case of old pharmacopeia analysis, the ability to accurately model and manipulate named entities is very important. Historical texts often contain a wide range of named entities that

must be extracted, linked, and analyzed to extract insights. Effective data modeling enables researchers to organize these entities in a manner that facilitates efficient querying and analysis. In this section, we aim to provide an overview of what has been done in representation and manipulation of named entities through exploring databases, as well as discussing other useful implemented models such as Knowledge Graphs and ontologies, all of which that have been useful means towards the end of querying structured data.

2.3.2 Databases

Databases are important in modern information systems by providing a structured way to store, retrieve, and manage data. They are useful for organizing large volumes of information and enabling efficient data manipulation. The concept of databases has evolved since their inception, reflecting the growing complexity and scale of data in various applications. Early database systems focused on simple, hierarchical structures, but modern databases support a wide range of data models and architectures, each suited to different needs and use cases. Among the most prominent types are relational databases and graph databases, both of which have distinct characteristics.

Relational Databases

Relational databases were introduced by Codd in 1970 and marked a step forward, as mentioned, in the field of database management. The relational model organizes data into tables linked among themselves, which are also known as relations. Each table is composed of rows and columns, where rows represent individual records, and columns represent the attributes of these records [89].This tabular structure allows for straightforward data organization and retrieval.

The relational model is based on main principles, including the ACID properties and normalization. ACID is an acronym that stands for Atomicity, Consistency, Isolation, and Durability, which are necessary for ensuring reliable transactions in a database [91]. Atomicity ensures that all operations within a transaction are completed; if any operation fails, the entire transaction fails. Consistency guarantees that a transaction brings the database from one valid state to another, adhering to all predefined rules and constraints. Isolation ensures that the operations of one transaction are invisible to other transactions until the transaction is completed which prevents concurrent transaction conflicts. Durability means that once a transaction is committed it remains committed even if the system crashes.

Normalization, in the context of relational database design, is the method of organizing data to reduce redundancy. This typically involves splitting a database into multiple tables and establishing relationships between them. The goal is to isolate data so that any additions, deletions, or modifications to a field can be executed in a single table and then reflected across the entire database through the defined relationships. These normal forms provide criteria to assess a table's susceptibility to logical inconsistencies and anomalies. The higher the normal form a table conforms to, the less prone it is to these issues. According to Demba (2013), normalization often necessitates creating additional tables, which some designers initially find challenging and cumbersome, despite the robustness of this kind of databases [92].

Relational Databases have been useful is many areas of study. In the context of botanical and medicinal plant research, which is relevant to our work on pharmacopeias that primarily describe plants for the composition of remedies, this has been especially impactful due to the need of structuring large amounts of data. For instance, the work of Allen [93] in 1993 focuses on the development and utility of botanical databases, specifically TROPICOS, managed by the Missouri Botanical Garden. The TROPICOS database, launched in 1983, serves as an extensive repository for plant names and associated data. This relational database is structured to facilitate a wide range of botanical research activities, from herbarium management to phylogenetic and ecological studies. The relational structure of TROPICOS allows for data querying and retrieval. Researchers can analyze plant species within a structured framework, examining phylogenetic relationships and geographical distributions. This capability is beneficial for large-scale projects like the Flora of North America (FNA), where the database supports analyses of plant species across the continent. Additionally, this work showed the integration of geographic information systems (GIS) with the botanical database which made possible the generation of maps that illustrate relationships between plant development and environmental factors such as climate change. This integration shows the potential versatility of relational databases in managing and analyzing complex datasets.

Similarly, Manhã et al. [94] introduce the PLANT database, a relational database designed to organize bibliographic information on medicinal plants in Brazil. The database categorizes data based on themes, taxonomic information, chemical substances, and pharmacological activities. By utilizing relational database structures, PLANT facilitates the efficient organization and retrieval of multidisciplinary information. The construction of the PLANT database involved the collection of literature from 22 periodicals, indexing articles based on predefined keywords. This systematic approach ensures that researchers can access relevant studies on medicinal

plants quickly. The relational nature of the database supports complex queries that allow its user to investigate connections between different plant species and their medicinal properties. PLANT also contributes to the preservation and dissemination of traditional knowledge about Brazilian medicinal plants. This aspect is interesting for ethnopharmacological research and the development of new medical drugs, which meets our own aim of uncovering useful information from Abbasid pharmacopeias.

Another work in this area is Syed and Khan [95] which describes the development of the Saudi Herbal Plants Information System (SHPIS), which uses MySQL, an open-source relational database management system (RDBMS), to store and manage data on medicinal plants. The SHPIS database catalogs 120 medicinal plant varieties, providing details on their local names, scientific names, medicinal uses, and parts used. The SHPIS web portal, built with Hypertext Preprocessor (PHP), offers an interactive platform for researchers and the public. It supports various query options, allowing users to search for plants based on different criteria such as family name, scientific name, and traditional usage. SHPIS also includes features for data submission and curation to aid researchers in their contribution to new information and updates. This collaborative aspect ensures the database's growth and evolution with new discoveries and insights.

Additionally, the work of Connelly et al. [4] studies the transformation of medieval medical texts into contextualized electronic databases. The study focuses on the 15th-century Lylye of Medicynes, a Middle English translation of Bernard of Gordon's Lilium medicinae. By converting this text into a relational database, the authors were able to analyze patterns in ingredient selection and usage. The database structure facilitated the organization of 3,548 ingredients used in 360 recipes to treat 124 unique diseases. This database enabled the researchers to conduct community detection analyses in order to identify core combinations of ingredients frequently used together to treat specific symptoms (Figure 15).

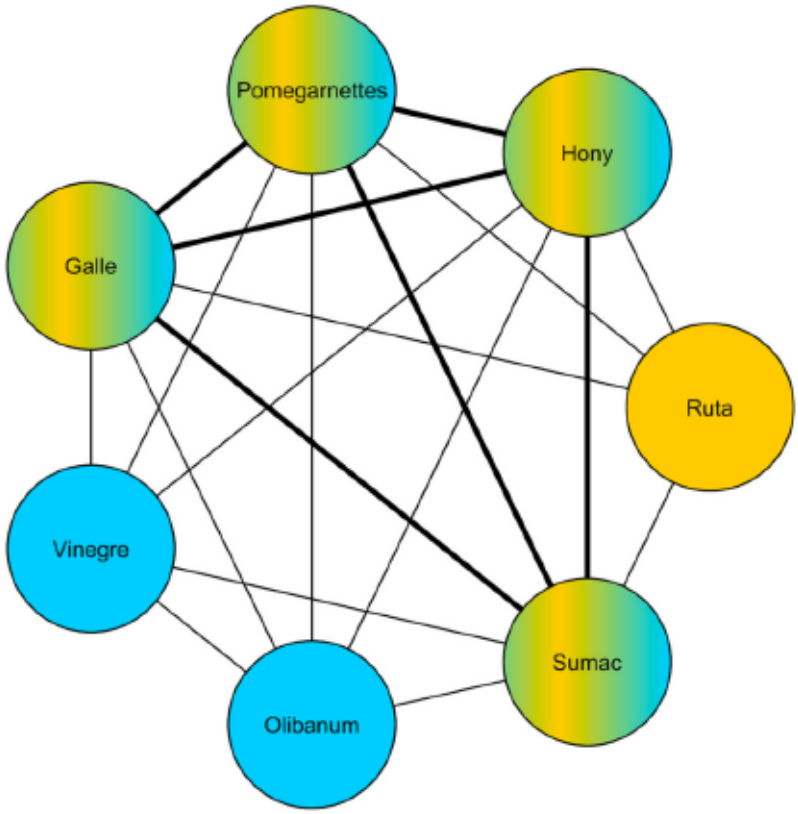


Figure 15 - Example of an ingredient network. The nodes in yellow are ingredients of a recipe for the treatment of fistula in lacrimali and the ones in blue are ingredients of a recipe for the treatment of pascionibus oris. The ones that are found in both recipes are colored both colors while thick link lines join pairs of ingredients that appear in both recipes, from [4].

This analysis revealed hierarchical communities of ingredients, reflecting patterns in medieval medical practices. Indeed, the ability to query and analyze these historical data sets using relational databases shows the potential of this approach in pharmacological research.

Graph Databases

Regarding Graph Databases, they represent a shift in database modeling and focus on the relationships between data points instead of organizing data into tabular structures. Introduced fore recently than their relational counterpart, in the mid-2000s, Graph Databases have gained popularity because of their ability to model and query interconnected data in an efficient way. Instead of using tables, Graph Databases use graph structures with nodes, edges, and properties to represent and store data [90], [96]. In a graph database, nodes represent entities, such as people, places, or objects, and edges represent the relationships between these entities. Each node and edge can have associated properties, which are key-value pairs that store additional information about the entities and relationships. This is useful for data that is qualitative and

relational, as its structure allows for the intuitive representation of interconnected data and makes easier the efficient traversal and querying of these connections.

One of the advantages of graph databases is their performance in querying complex relationships. Traditional relational databases often require multiple joins to traverse relationships, which can be computationally expensive. In contrast to this, graph databases are good at these operations due to their native graph traversal algorithms, which can traverse relationships directly through edges. Indeed, this can result in performance gains for certain types of queries.

Further, Graph Databases also support ACID properties which we described as being properties that ensure reliable transactions. As the demand for managing and analyzing complex, interconnected data continues to grow, graph databases are becoming increasingly used across different fields, and this is relevant to our work on old pharmacopeia analysis. Their ability to model real-world systems more naturally than traditional relational databases makes them a useful tool. Additionally, graph databases are flexible and can more easily accommodate changes in the data model. This is especially useful in environments where the schema may evolve over time. In contrast to that, modifying the schema of a relational database can be a complex and time-consuming process when dealing with large volumes of data.

In the context of cultural heritage, Spadini et al. [97] describes the integration of graph data models and semantic web technologies. Their work describes the application of graph database technology, among other tools. Through using graph databases, they represent complex relationships between artifacts, historical contexts, and metadata, which allows for detailed analysis and discovery of hidden patterns. The study demonstrates how graph databases can link artifacts to their historical usage, geographical origins, and related literature and provides a wide view of cultural assets. This integration showcases the potential of graph databases to improve data management and analysis in the cultural heritage sector.

In the context of plant-based data, Singh et al. [98] introduced the DISPEL database, which makes use of Neo4j, a graph database management system, to manage and visualize relationships between medicinal plants and the diseases they cure. Figure 16 illustrates the overall strategy of the creation of the database.

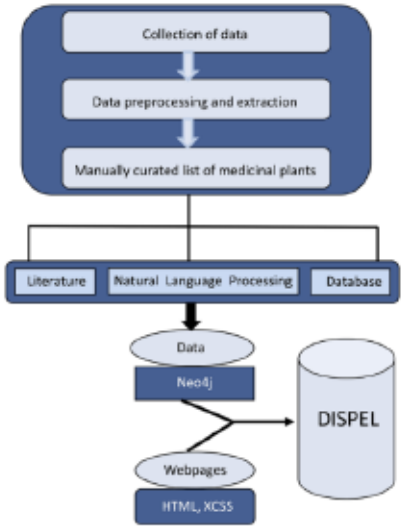


Figure 16 - Representation of the employed strategy in the creation of the DISPEL database, from [98].

The DISPEL database hosts approximately 60,000 plant-disease linkages, encompassing around 5,500 medicinal plants and 1,000 diseases. This extensive dataset is organized into a network graph, where nodes represent medicinal plants and diseases, and edges represent the therapeutic relationships between them. The graph-based representation allows users to perform complex queries to identify the most effective medicinal plants for specific diseases. The interactive visualization capabilities of the database allows researchers to study the network graph and understand the relationships visually. This approach aids in drug discovery and improves the understanding of traditional medicinal knowledge.

Aditionally, the work of Chelazzi and Bonzano [99] discusses the application of integrative big data approaches, including the use of graph databases, in archaeological data related to the ancient Mediterranean region. They shed light onto how graph databases can connect disparate data points and facilitate complex queries across large datasets. The authors show the importance of graph databases in handling heterogeneous data sources. By representing data as nodes and edges, graph databases efficiently integrate information from different domains. This capability is useful in interdisciplinary research, where data from various fields need to be analyzed together to address complex scientific questions. Graph databases allows for efficient graph traversals and pattern matching, essential for identifying relationships and patterns not immediately apparent in traditional tabular data.

In the case of agricultural data, Abad-Navarro et al. [100] present a workflow for generating scientific literature knowledge graphs in the agriculture domain and translating them into property graphs implemented in Neo4j. The study illustrates how graph databases can improve

literature searches and data integration in scientific research. The authors describe a pipeline that converts scientific publications into RDF, enriches the content with semantic annotations, and populates a property graph with this information according to the model shown in Figure 17.

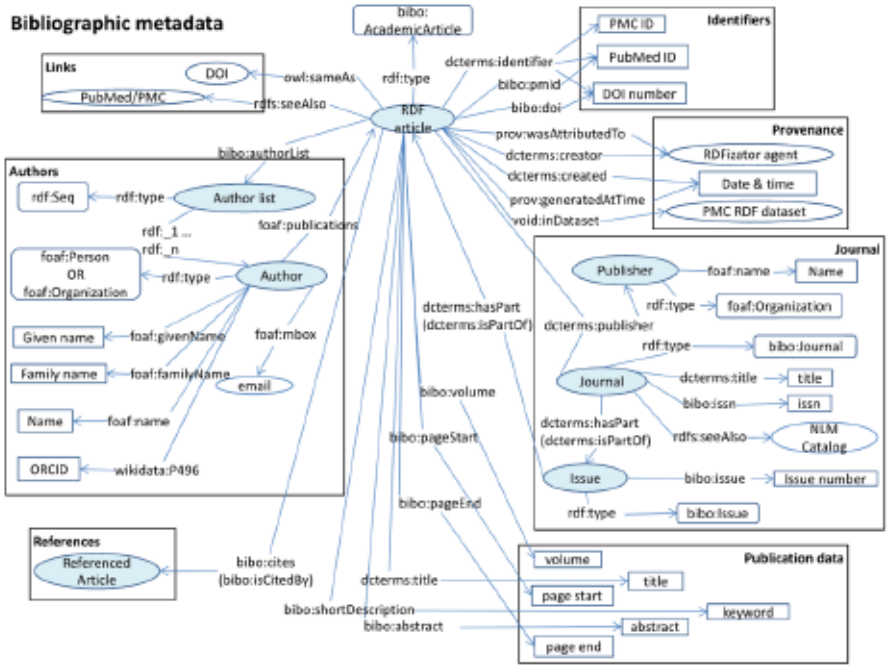


Figure 17- The model that is used for describing article metadata, from [100].

The resulting graph includes metadata, article content, and domain-specific annotations, which provides an interconnected dataset that can be queried using Neo4j's Cypher query language. The knowledge graph generated contains information about 127 agriculture-related articles which indicates the potential of graph databases to manage and analyze corpora of scientific literature. The paper describes several use cases, including literature search, article similarity analysis, and semantic clustering and demonstrates the advantages of using graph databases in scientific research.

Table 1 is an overview of the different databases, graph and relational, that we have reviewed

Database	Database Type	Purpose/Focus	Geographic Focus	Key Feature
TROPICOS (Allen [93])	Relational Database	Botanical research	Global	GIS integration

PLANT (Manhã et al. [94])	Relational Database	Medicinal plants	Brazil	retrieval of multidisciplinary information
SHPIS (Syed and Khan [95])	Relational Database	Medicinal plants	Saudi Arabia	Interactive web portal, collaborative data submission
Lylye of Medicynes DB (Connelly et al. [4])	Relational Database	Medieval medical texts	England	Analysis of ingredient patterns
DISPEL (Singh et al. [98])	Graph Database	Medicinal plants and diseases	Global	Interactive visualization, complex querying regarding diseases
Cultural Heritage DB (Spadini et al. [97])	Graph Database	Scholarly and cultural heritage data	Global	RDF integration
Archaeological DB (Chelazzi and Bonzano [99])	Graph Database	Archaeological data	Mediterranean region	Integration of disparate data, complex queries
Agricultural DB (Abad-Navarro et al. [100])	Graph Database	Agricultural scientific literature	Global	Literature search, article similarity analysis

Table 1 - Summary of the reviewed databases in the literature

These works collectively shed light onto the impact of graph databases across various domains. Indeed, graph databases offer robust solutions for managing, querying, and visualizing complex, interconnected data. Their ability to model and analyze relationships directly and efficiently makes them relevant tools in data management.

In conclusion of this section, we see that both relational and graph databases offer distinct advantages in the representation and manipulation of named entities. Relational databases provide a structured and reliable means of storing and querying data which makes them suitable for applications where data integrity and consistency are the most important aspect. Graph

databases, on the other hand, are suited in scenarios where complex relationships are involved, offering greater flexibility and efficiency in traversing and analyzing interconnected data. The choice between relational and graph databases depends on the specific requirements of the application, with each model offering unique strengths that can be used to achieve optimal results in within the chain of text mining tasks.

2.3.3 The Use of Ontologies

Ontologies serve as formal representations of knowledge by defining a set of concepts and the relationships among them within a specific domain. According to Gruber, an ontology is "a specification of a conceptualization" [101]. Through this definition, we can see the prescribed importance of ontologies in structuring domain-specific knowledge in a formal manner which would in turn make them relevant for fields requiring complex data integration and semantic analysis. Since text mining is the process of extracting meaningful information and patterns from large text datasets, the integration of ontologies into text mining tasks can benefit this process because of the provision of a structured vocabulary as well as explicit relationships between concepts. The relevance of ontologies to text mining can be seen in a few areas such as knowledge representation, where ontologies provide a standardized way to represent domain knowledge in a machine-readable format that facilitates automated reasoning and knowledge discovery. Additionally, it can be seen when it comes semantic enrichment, because by linking text data to ontological concepts, text mining tools can better take into account the context and semantics of the data. Further, they facilitate interoperability and data integration because they enable the integration of heterogeneous data sources through providing a common framework for describing data. This is very important in medical research for example, where data from different studies and resources need to be combined and analyzed cohesively.

The development of the Arabidopsis Thaliana ontology [102] is an example of the usage of ontologies in plant science to model growth and developmental stages. For instance, this described ontology takes into account detailed information about the morphology and development of Arabidopsis Thaliana which would allow researchers to query and reason about the plant's biology. Tools like Apache Jena Fuseki and Protégé were used to construct and query the ontology, demonstrating the practical implementation of ontologies in data integration.

Additionally, the Plant Phenology Ontology (PPO) [103] exemplifies the use of ontologies to integrate large-scale phenological data. The PPO makes use of the Plant Ontology (PO) to

provide a consistent vocabulary for describing plant developmental stages, facilitating the aggregation and analysis of phenological data across different studies and data sources. This ontology is relevant for studying the impacts of climate change on plant phenology by enabling the comparison and integration of data recorded at different levels of precision.

In another relevant work on plant-based data, Knomana [104] is a knowledge-based system designed to organize and manage knowledge on the different uses of plants, specifically in pest control. The ontology within Knomana facilitates the systematic representation of relationships between plants, target organisms, and protected systems. The system employs formal concept analysis (FCA) and relational concept analysis (RCA) to navigate the knowledge base and allows for the extraction of relevant knowledge patterns and supporting decision-making in various health domains, such as animal health and environmental health.

Indeed, ontologies are very useful in enhancing the capabilities of text mining through structured knowledge representation, semantic enrichment, and data integration. However, the increasing complexity of maintaining and querying large ontologies as data volumes grow necessitates the development of efficient algorithms and scalable infrastructure to manage the data effectively. Additionally, the challenge of integrating data from diverse sources with varying ontological commitments shows the need for universal standards and mappings to ensure efficient data integration, which may prove to be a difficult task when it comes to the analysis of ancient medical pharmacopeias where the data is less structured than modern medical documents. It is worthwhile, in this work, to study the possibility of mapping the various entities present in ancient pharmacopeias to real-world objects, all while making sure the used resources are consistent to ensure their effectiveness and their relevance to domain-specific requirements.

2.3.4 Insight Extraction: The Case of Formal Concept Analysis

The extraction of insights from structured data, especially in the context of historical medical documents and pharmacopeias proves to be a worthwhile task. The goal of insight extraction is to discern patterns and relationships among various entities present within these documents to inform contemporary medical research. This review focuses on Formal Concept Analysis (FCA) as a robust method for extracting such insights.

FCA is a mathematical framework used to analyze and represent data in terms of concepts and their hierarchical relationships. Originating from lattice theory, FCA provides a structured way to discover and represent structures within data sets. This method is effective for text mining

and information retrieval due to its ability to elicit context and organize data into a formal structure called a Galois lattice or concept lattice [105], [106].

FCA operates by identifying formal concepts within a given dataset, where each concept comprises a set of objects (e.g., ingredients) and a set of attributes (e.g., symptoms treated by these ingredients). These concepts are then organized into a lattice structure, which reveals the hierarchical relationships between them. The lattice structure allows for the efficient querying and retrieval of information which makes of it a useful tool for researchers and domain experts. FCA has been successfully applied in numerous domains, demonstrating its versatility and effectiveness. For instance, Braud et al. [107] presented a work in which FCA is used to discover co-occurrences of pharmacopeia-based ingredients and indicated the suitability of this method for insight extraction based on the need of experts. For example, Figure 18 is a lattice that identifies which ingredients commonly co-occur with "celery seeds" across different sources. The top concept represents "celery seeds" while subsequent concepts show the co-occurring ingredients under various books, we see that "fennel seeds" most frequently appears with "celery seeds" (7 times) which may be a beginning of an insight when it comes to analyzing ingredient co-occurrences by specialists. Figure 19 shows an example from the same work regarding implication rules, are "if-then" statements indicating that if a set of attributes A is present, then another set B must also be present [108]. In this case, the figure shows implication rules where the presence of one set of attributes (e.g., "Celery seeds") consistently implies the presence of another (e.g., "Fennel seeds"). The support values indicate how frequently these relationships occur.

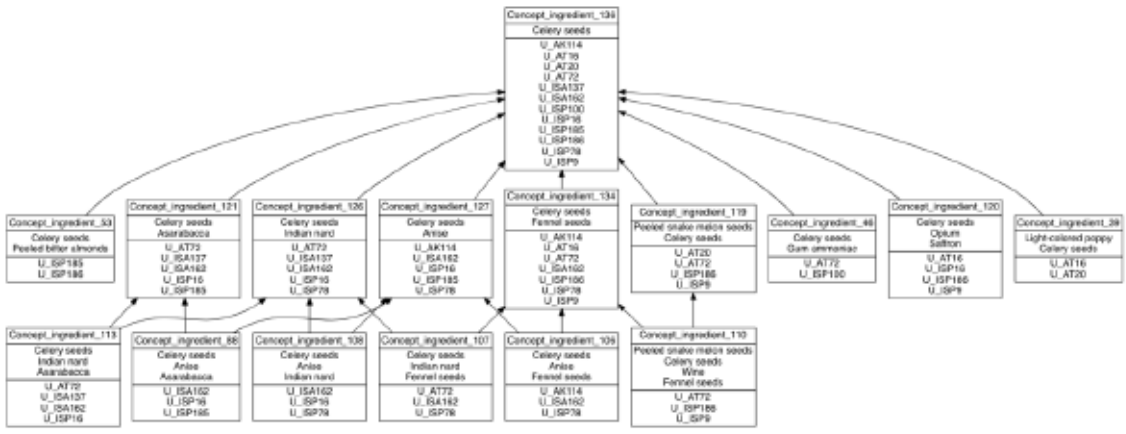


Figure 18 - Example of a lattice showing which ingredients co-occur with celery seeds, from [107]

Rule	Support
Celery seeds → Fennel seeds	7
Opium → Saffron	6
Asarabacca, Indian nard → Celery seeds	4

Figure 19 - Example of implication rule, from [107]

This work indeed demonstrated the need of enriching the databases that were populated out of pharmacopeia data to be able to perform more complex analyses. Further, Allard et al. [109] employed FCA to discover functional dependencies and association rules in a lattice of OLAP views which also demonstrated its utility in data analysis and knowledge discovery.

In a different work, Braud et al. [110] also utilized a lattice-based query system, this time to assess the quality of hydro-ecosystems, which indicates FCA's ability to manage and interpret complex ecological data. These examples underline FCA's potential to handle diverse datasets and extract meaningful insights.

In the context of analyzing old pharmacopeias, FCA can be used to answer several relevant questions. For instance, identifying the most common ingredients, understanding which ingredients are frequently used together, and determining the associations between ingredients and specific symptoms or organs. This approach is exemplified in the work of Connelly et al. [4], who applied network analysis techniques, different from FCA, to a medieval medical text to reveal patterns in ingredient choice that reflect their biological activity against infectious agents. FCA can extend this analysis by organizing these patterns into a lattice structure and aid in the identification of broader trends and relationships. This method's efficacy is also demonstrated by Silvie et al. [104], who developed a knowledge-based system to identify botanical extracts for plant health in Sub-Saharan Africa using FCA.

Thus, through structuring information into a lattice, researchers may gain an omniscient view of the data which would facilitate spotting correlations and potential areas for further research.

Chapter 3: Building and Assessing a Named Entity Recognition Resource for Ancient Pharmacopeias

3.1 Introduction

The primary goals of this chapter are twofold. First, to identify the best transformer-based model for performing NER on English translations of Arabic pharmacopeias. Second, to assess the generalizability of this model across different manuscripts and translation styles.

The first phase of the study involved comparing the performance of several transformer-based models, on a single annotated manuscript. The goal was to determine which model achieved the highest accuracy in recognizing and classifying entities within the text. This phase was necessary for identifying the most effective model architecture and pre-training strategy for the specific domain of historical medical texts. Each model was fine-tuned on the annotated dataset, and their performance was evaluated based on precision, recall, and F1-score. DeBERTaV3 emerged as the best-performing model, achieving the highest F1-score, and was selected for further experiments.

The second phase of the study expanded the analysis to include two additional annotated manuscripts. The selected model from the first phase (DeBERTaV3) was trained on various combinations of manuscripts to evaluate its performance and generalizability across different texts and translation styles. This phase aimed to determine how well the model could adapt to new texts and identify factors that influence its performance, such as the diversity of the training data and the differences in linguistic and stylistic features among the manuscripts.

3.2 Data and Annotations

The study focused on three manuscripts that are in our possession and studied by our historian colleagues, each representing important medical texts from the medieval Arabic world. These manuscripts, translated into English, provided the primary data sources for the NER tasks.

- Sabur ibn Sahl's Dispensatory in the Recension of the Adudi Hospital (9th century Baghdad), translated by Oliver Kahl [111] (pharmacopeia of 37k words).
- The Dispensatory of Ibn at-Tilmīdh (12th century Baghdad), translated by Oliver Kahl [112] (pharmacopeia of 40k words).
- Ibn al-Jazzar's Provision for the Traveler and Nourishment for the Sedentary (10th century, Kairouan) - Book 7, translated by Gerrit Bos [113] (pharmacopeia of 16k words).

The first manuscript we worked on was Oliver Kahl's English translation of the Dispensatory in the Recension of the Adudi Hospital, written by Sabur ibn Sahl in the 9th century. The dispensatory, attributed to Sabur ibn Sahl, a prominent Persian Christian physician and pharmacologist operating at the Academy of Gondishapur before moving to Baghdad, showcases the pharmacological practices of the time. Sabur ibn Sahl's work, through its recension under the auspices of the Adudi Hospital, demonstrates a systematic approach to drug composition and therapeutic applications. We chose this manuscript for its well-structured style of writing, as the text has also been edited by the translator-historian who is the chooser of the authoritative copy of the old manuscript. The choice of a well-structured corpus was made because our hypothesis that is might positively affect the accuracy of data extraction, as relevant entities and relationships are more easily identified. A clear and organized manuscript reduces the complexity of preprocessing steps, saving time and resources in the overall work.

The first half of the document contains the text in its original Arabic whereas the second half is the English Translation by Oliver Kahl, the latter half being the focus of our work. The corpus is divided into chapters based on drug categories or therapeutic applications, such as pastilles, lohochs, beverages, oils, cataplasms, enemas, powders, and collyria. Each entry within the chapters provides information on the preparation, dosages, and intended therapeutic use of the compounds. In total, the corpus describes 292 remedies encompassing a wide array of substances and ingredients from various geographical origins including vegetable, animal, mineral, and, occasionally, human substances. Similarly, and by the same translator, we also incorporate The Dispensatory of Ibn at-Tilmīdh that was composed in 12th century Baghdad. This physician also worked in the Adudi hospital and where his predecessor of 3 centuries Sabur Ibn Sahl worked. Indeed, Ibn at-Tilmīdh built upon Sabur ibn Sahl's pharmacopeia and wrote an extensive pharmacopeia of 424 remedies. We choose this manuscript for its structural similarity to the previous manuscript as well as its translation work done by the same person, albeit originating from a different century, as this would impact the study of the generalizability of the NER method.

Within each chapter, individual entries provide detailed information for each described remedy. These entries typically include:

- Name and Description: the number, name of the remedy and a brief description of its intended use or therapeutic properties and the symptoms or pathologies it aims to treat.
- Ingredients: a list of components used in the preparation, often with precise quantities or proportions. This includes a diverse array of substances.

- Preparation instructions: steps or actions to be taken for preparing the pharmaceutical compound.
- Application and Dosage: guidelines for the administration of the medicine as well as dosages.

Afterwards, we chose Book 7 of Provision for the Traveler and Nourishment for the Sedentary, by Ibn al-Jazzar and translated by a different scholar than the first two documents, Gerrit Bos. This 10th-century manuscript by Ibn al-Jazzar, a prominent physician from Kairouan, focuses on practical medical advice for travelers and sedentary individuals. This pharmacopeia covers a wide range of remedies and preventive measures for various ailments. This pharmacopeia has a more discussion-like structure where discussion elements are intertwined within the recipe description after presenting the symptoms of a disease and numbering paragraphs wherever it was deemed fit in the editing. Its description of remedies differentiates itself from the first two pharmacopeias in that it introduces diseases as general chapters and then proceeds to discuss the many different methods of treating those diseases in numbered paragraphs. We choose this manuscript to complete our corpus for this study for the diversity it brings in terms of difference of structure, century and location of origin and translator.

Editing old manuscripts is indeed a minute task of the historian that requires making specific choices regarding the interpretation, presentation and arrangement of handwritten text elements which can directly influence the structure of the corpus. Figure 20 showcases how remedies from the 3 manuscripts present themselves in the translated and edited version in English.

Ibn at-Tilmid (Oliver Kahl):
(41)
The spikenard pastilles
for (the treatment of) an inveterate tumour in the stomach
Citronella blades, cassia, roses, rhubarb, lemon grass, and Indian spikenard three dirham of each; saffron, anise, alecost, and black pepper one dirham of each; bdellium africanum three dirham; mastic two dirham; ammoniacum one dirham. (This) is formed into pastilles, (and) a potion (may be made by using) one mitqal (of it) every day with wine boiled down to one quarter.

Sabur Ibn Sahl (Oliver Kahl):
[93]
A cataplasm for (the treatment of) swollen glands
Take pure bdellium mukul, Yemenite alum, mastic, and pomegranate flowers in equal (parts). (This) is pounded, kneaded with fresh myrtle-water, and applied as a cataplasm.

Ibn al-Jazzar (Gerrit Bos):
Chapter 18: On baraṣ and bahaq
(1) Baraṣ and bahaq have the same origin and should be treated in the same way. For baraṣ originates from the corruption of the blood with which the skin of the body feeds itself, while bahaq originates from the corruption of the blood with which the visible layer of the skin of the body feeds itself, but the part beneath it is not affected. This is the difference between baraṣ and bahaq. [...] In order to treat baraṣ and white bahaq one should administer the patient a decoction of epithyme and agaric with hiera picra which he should take in the spring season. His body should be purged with hieras containing pulp of colocynth, such as the great hieras, the Logadius, the Theodoretus, and the great stomaticum and the like.

Figure 20 - Different styles of translation and editing from the 3 studied pharmacopeias

After this task of corpus selection, the preparation of the data involved several steps to ensure the texts were suitable for annotation and subsequent model training.

Text Conversion. The original PDF versions of the manuscripts were converted into plain text format using the PdfToText library[12]. This conversion was necessary to facilitate further text processing and annotation. The tool extracted the textual content while preserving the structure and chronological order of the elements of the documents.

Cleaning and Normalization. The converted texts were cleaned to remove non-essential elements such as discursive footnotes, introductions, and prefaces. This step was important to eliminate noise and ensure that only relevant content was processed, that of the pharmacopeia. The remaining text was then normalized to ensure consistency in formatting, which is important for accurate tokenization and annotation. Normalization involved standardizing punctuation, correcting OCR errors, and ensuring uniform text formatting.

Tokenization. The normalized text was tokenized using the NLTK library [114]. Tokenization involves breaking the text down into individual tokens (words or phrases) that can be annotated with entity labels. This process is fundamental to preparing the text for NER, as it defines the units of text that the model will analyze.

The annotation process was a central component of the study, as it is the basis upon which the NER task can be performed. Through the expertise of domain specialists including historians and pharmacognosists, and to ensure the accuracy and relevance of the annotations, it was decided that the following entity types were to be annotated:

- Ingredients (ING): Substances used in the preparation of medicinal remedies.
- Symptoms (SYM): Descriptions of medical symptoms and conditions.
- Organs (ORG): References to parts of the human body.
- Preparation Types (TYPE): Methods and forms of preparation for medicinal remedies.

As our NER task lies within a pipeline of Text Mining tasks which aims to extract insights from old pharmacopeias which answer questions such as "what ingredients (ING) occur frequently together to treat fevers (SYM)", "what forms (TYPE) of remedies are most commonly used to treat the skin (ORG)?". As such, the decision was made to focus on those 4 entity types, leaving other potential tags to future studies (quantities, units of measurement, verbs of action, etc.)

[12] https://pypi.org/project/pdftotext/

The annotation was conducted using the IOB2 format, a standard format for tagging tokens in NER tasks. In this format, the beginning of an entity is marked with a "B-" prefix, the inside of an entity with an "I-" prefix, and tokens outside any entity with an "O" label. This structured format helps in precisely identifying the boundaries and categories of entities within the text.

The annotation process involved several stages. The initial annotation was performed by domain experts, including a computational linguist, historians, and pharmacognosists manually. The computational linguist is the main annotator of the corpus and the pharmacognosists and historians were available to provide their expertise and advise the annotator in cases of doubt. Each token was examined and tagged with the appropriate entity label based on its context within the sentence. Afterwards, the annotations were reviewed by said experts to ensure consistency and accuracy. This stage involved resolving any ambiguities and discrepancies in the annotations. Indeed, this verification amounted to final validation to confirm that all entities were correctly labeled, and that the dataset was ready for model training. This step ensured the reliability of the annotations and the overall quality of the dataset.

The annotated dataset was analyzed to determine the distribution of entity types across the three manuscripts. The tag counts for each entity type are summarized in the following table:

Manuscript	Ingredient Tags	Symptom Tags	Organ Tags	Type Tags	Total Tokens	Annotation team	Time spent on annotation
Ibn al-Jazzar	2,874	903	113	146	16,278	Computational Linguist with Historian and Pharmacognosist expertise and support	1 month
Ibn at-Tilmīdh	8,052	1,365	159	771	40,330		1 month
Sabur ibn Sahl	5,789	1,504	177	396	36,960		1 month

Table 2 - Tag counts for each annotated pharmacopeia

The distribution of tags reflects the content and focus of each manuscript, with Ibn at-Tilmīdh's manuscript containing the highest number of annotated entities. These tag counts provide a basis for evaluating the performance of the NER models and understanding their ability to

generalize across different texts and entity types. A closer look at the tag counts reveals several insights into the nature of the manuscripts:

Ingredients (ING). This entity type had the highest frequency across all manuscripts. Indeed, ingredients are the basis for pharmacopeias, as they describe the diverse plants, minerals and animal-based substances that are used to ease symptoms and cure diseases. They have been taken into account as having the ING tag as a whole, even when they have their descriptive transformation or parts mentioned (such as "grilled apple seeds" which would be considered as ING as a whole)

Symptoms (SYM). Symptoms were the second most frequent entity type. The manuscripts place much emphasis on discussing and treating medical conditions as correlated to the mentioned ingredients.

Organs (ORG). This entity type had the lowest frequency, suggesting that while references to body parts were important, they were less common than descriptions of ingredients and symptoms. Organs occur less commonly as standalone mentions in the text for them to be annotated as such. Indeed, they often occur within a symptom occurrence such as "abdominal disorders". In such cases, this entity is annotated as a symptom. While the adjective "abdominal" may indicate a mention of an organ nested within a symptom, we choose not to perform nested named entity annotation for this study to avoid additional complexity for the NER task because of the relatively small size of the training data.

Preparation Types (TYPE): The tags for preparation types were relatively consistent across the manuscripts in proportion to their respective sizes, reflecting a tendency of the standardizing forms of preparation for remedies in Abbasid pharmacopeias.

3.3 Methodology and Experiments

The methodology and experimental setup are designed to systematically evaluate the performance and generalizability of various transformer-based models for Named Entity Recognition on historical medical texts. This section outlines the experimental design, training setups, and the results obtained from these experiments.

The first phase of the study focused on comparing the performance of several transformer-based models on a single annotated manuscript: Sabur ibn Sahl's Dispensatory in the Recension of the Adudi Hospital. The models that we compare are the following:

- BERT: A transformer-based model pre-trained on a large corpus of English text from Wikipedia and BookCorpus [45].
- RoBERTa: An optimized version of BERT with more training data and longer training duration [115].
- XLM-R: A multilingual version of RoBERTa, pre-trained on a large corpus of text in multiple languages [116].
- BioBERT: A version of BERT pre-trained on biomedical text, including abstracts and full-text articles from PubMed [16].
- DeBERTaV3: A transformer model with an improved attention mechanism and enhanced pre-training strategy [11].

Each model was fine-tuned on the annotated dataset, and their performance was evaluated based on precision, recall, and F1-score. This comparative study was the main topic of our work in a publication published in 2023 [21].

To determine the best model and the best set of hyper-parameters, we ran all the experiments with a 5-split cross-validation set without any shuffling, since sentence-level shuffling may result in data leakage between training and validation data. Table 3 shows the different values of the set of hyper-parameters that were used. Each model is trained on the combination of hyper-parameters for a maximum of 10 epochs, i.e. a total of 480 runs over a 33-hour period, and only the results of the best epoch are considered. The experiments were run using an RTX 3080Ti with 12GB RAM with the HuggingFace Transformers library [117].

Hyper-parameters	Values
Max Sequence Length.	256
Batch Size	32
Learning Rate	{4,5,6,7}e-5
Warm-up	{0,0.1}
Scheduler	{linear, cosinus}

Table 3 - Hyper-parameters used for fine-tuning

Table 4 shows the average F1 score, precision and recall over the 5 splits from the best set of hyper-parameters. DeBERTaV3 achieves the best performance of all models, but also the least variable, outperforming BioBERT, which was trained on a corpus from the medical domain. We also note that the XLM-R multilingual model performed less well than its English counterpart RoBERTa. Finally, the original BERT model gave the lowest score.

Model	Precision	Recall	F1
XLM-R	83.36 ± 2.53	84.97 ± 4.35	84.12 ± 2.92
BERT	83.09 ± 1.92	86.19 ± 5.40	84.26 ± 3.33
BioBERTv1.2	83.47 ± 1.52	85.93 ± 4.15	84.66 ± 2.46
RoBERTa	84.78 ± 2.34	86.39 ± 3.63	85.56 ± 2.14
DeBERTaV3	**85.78 ± 1.15**	**87.09 ± 2.46**	**86.03 ± 1.55**

Table 4 - Mean value and standard deviation of the 5 iterations for the best set of hyper-parameters

When tuning the hyper-parameters, DeBERTaV3 consistently outperformed all other models, whatever the set of hyper-parameter values, showing the advantage of DeBERTaV3's disentangled attention mechanism, which makes it the suitable choice for the NER task that we expand upon to continue the study of generalizability of NER for the goal of testing transferability to a new translator and author, we have categorized our experiments into three distinct groups to systematically evaluate the performance of our NER model in three different scenarios[13]:

Single Source Manuscript Training

In the single source manuscript training setup, the DeBERTaV3 model was trained on the annotated dataset of one manuscript and tested on the other two manuscripts. This experiment aimed to evaluate the model's ability to generalize from one manuscript to another with different content and translation styles. The annotated dataset from one manuscript was used for training, while the annotated datasets from the other two manuscripts were reserved for testing. The DeBERTaV3 model was fine-tuned on the training dataset using a batch size of 32, a maximum sequence length of 256, and a learning rate of 5e-5 for 10 epochs. Five different random seeds were used to mitigate the impact of seed-specific variations. The trained model

[13] Main topic of an accepted paper: K. El Haff, W. Antoun, A. Braud, F. Le Ber, V. Pitchon, "Building and Assessing a Named Entity Recognition Resource for Ancient Pharmacopeias", in 27TH European Conference On Artificial Intelligence, Santiago de Compostela, October 2024

The fine-tuned models and experiments are openly available on Huggingface for the community via https://huggingface.co/karimelhaff/remed-ner

was evaluated on the test datasets using standard precision, recall, and F1-score as performance metrics.

Combining Two Manuscripts for Training

To improve our understanding the model's generalizability, the next set of experiments involved training the DeBERTaV3 model on a combination of two manuscripts and testing it on the third manuscript. This setup aimed to determine whether training on a more diverse dataset improves the model's performance on unseen texts. Annotated datasets from two manuscripts were combined for training, while the annotated dataset from the third manuscript was reserved for testing. The DeBERTaV3 model was fine-tuned on the combined training dataset using the same training parameters as in the single source manuscript training. This experiment helped to assess the impact of training on diverse datasets and provided insights into the model's ability to generalize across different translation styles and content.

Combining All Manuscripts for Training

The final experimental setup involved training the DeBERTaV3 model on the combined annotated datasets of all three manuscripts and evaluating its performance on each manuscript individually. This setup aimed to maximize the diversity of the training data and assess the model's overall performance. Annotated datasets from all three manuscripts were combined for training. Similarly, the DeBERTaV3 model was fine-tuned on the combined training dataset using the same training parameters as in the previous two experiments.

To ensure consistency in our evaluation, we used an 80/20\% train-test split for all manuscript annotations. This ensured that each experiment had a consistent evaluation protocol. We note that the 3 manuscripts are of different lengths (Ibn Jazzar being about half the size of Sabur Ibn Sahl and Ibn Tilmīdh, as shown in Table 2), and since it is needed to account for the variability in the training corpus size, in the first set of experiment, we limit the size of the resulting training corpus of each manuscript to match the smallest manuscript (Ibn al-Jazzar) which the resulted in that the training corpus consists of a split of the two other sources to match the size of Ibn al-Jazzar on one hand (Table 5) and a training corpus with the full sizes on the other hand (Table 6).

The results for experiments with the adjusted and full manuscript sizes (5-seed average) are summarized in the following tables:

Training set/Text set	Ibn al-Jazzar *(Gerrit Bos)*	Ibn at-Tilmīd *(Oliver Kahl)*	Sābūr ibn Sahl *(Oliver Kahl)*
Single Training Manuscript			
Ibn al-Jazzar	75.63 ± 1.80	78.25 ± 1.61	80.55 ± 0.97
Ibn at-Tilmīd	69.09 ± 1.73	82.34 ± 0.61	81.59 ± 0.89
Sābūr ibn Sahl	69.57 ± 2.28	81.35 ± 0.52	81.06 ± 0.97
Two Training Manuscripts			
Ibn al-Jazzar/ Ibn at-Tilmīd	76.31 ± 0.81	83.58 ± 0.39	82.14 ± 0.96
Ibn al-Jazzar/ Sābūr ibn Sahl	76.75 ± 1.33	82.04 ± 1.15	82.12 ± 1.05
Ibn at-Tilmīd/ Sābūr ibn Sahl	70.51 ± 2.86	81.70 ± 1.29	80.37 ± 0.85
Training On All Manuscripts			
All	75.91 ± 0.78	83.60 ± 0.61	83.21 ± 0.97

Table 5 - F1-scores for experiments with the equal training dataset size (5-seed avg.)

Training set/Text set	Ibn al-Jazzar *(Gerrit Bos)*	Ibn at-Tilmīd *(Oliver Kahl)*	Sābūr ibn Sahl *(Oliver Kahl)*
Single Training Manuscript			
Ibn al-Jazzar	75.63 ± 1.80	78.25 ± 1.61	80.55 ± 0.97
Ibn at-Tilmīd	74.69 ± 1.92	87.09 ± 0.54	84.15 ± 0.52
Sābūr ibn Sahl	74.68 ± 1.65	85.17 ± 0.42	85.15 ± 0.65
Two Training Manuscripts			
Ibn al-Jazzar/ Ibn at-Tilmīd	79.70 ± 0.88	87.54 ± 0.36	85.35 ± 0.46
Ibn al-Jazzar/ Sābūr ibn Sahl	79.93 ± 1.35	85.53 ± 0.65	86.95 ± 0.65
Ibn at-Tilmīd/ Sābūr ibn Sahl	76.63 ± 0.96	86.24 ± 0.26	86.12 ± 1.50
Training On All Manuscripts			
All	80.91 ± 0.85	86.51 ± 0.69	87.21 ± 0.62

Table 6 - F1-scores for experiments with the full manuscripts (5-seed avg.)

Indeed, these results provide insights about the performance and behavior of the model through the different experiments, in regard to effects of the different changes that were made to perform the experiments:

Effect of Original Manuscript Author

A main aspect of the study was evaluating the transferability of the NER model across manuscripts translated by different scholars. These experiments aimed to understand how well a model trained on data translated by one individual could perform on data translated by

another. The results from the single manuscript training experiments with adjusted sizes provided insights into this transferability. The impact of translator variance on model performance is present, but not drastically so, and hinting towards a good generalizability of the NER model. By training on manuscripts translated by Oliver Kahl (Ibn at-Tilmīdh and Sabur ibn Sahl) and testing on Gerrit Bos's translation of Ibn al-Jazzar, F1-scores of 69.09 for Ibn at-Tilmīdh) and 69.57 Sabur ibn Sahl were achieved, which demonstrates a notable yet not debilitating dip when compared to the intra-translator testing score of 75.63 for Ibn al-Jazzar when trained and tested on itself. On the other hand, training on Ibn al-Jazzar showed slightly better inter-translator performances (78.25 and 80.55 vs 75.63). Similarly, the F1-score for training on Ibn al-Jazzar and testing on Ibn at-Tilmīdh was 78.25, and the F1-score for training on Sabur ibn Sahl and testing on Ibn at-Tilmīdh was 81.35 which suggests a better performance when the manuscripts are of the same translator. Indeed, training only on manuscripts by Ibn at-Tilmīdh or Sabur ibn Sahl yields similar scores on their respective test sets, suggesting that the model's performance is not as much affected by the original manuscript author as when changing translators.

These results indicate that while the model could generalize reasonably well, there was a noticeable drop in performance when switching between translators. This effect may be due to the nature of Ibn al-Jazzar's pharmacopeia including more discussion elements and a different editing style and thus different entity distribution compared to the more succinct and structured style of Ibn at-Tilmīdh and Sabur ibn Sahl shared by the same translator, as mentioned in Section 3.2.

Effect of Mixing Training Manuscripts

Mixing training manuscripts was hypothesized to improve the model's robustness by exposing it to a wider variety of linguistic and contextual nuances. The results from combining two manuscripts for training supported this hypothesis through showing improved performance compared to single manuscript training. Mixing data from different manuscripts clearly offsets the translator effect and boosts the performance overall.

Training on a combination of Ibn al-Jazzar and Ibn at-Tilmīdh or Ibn al-Jazzar and Sabur ibn Sahl resulted in higher scores on manuscripts from a different translator, compared to training solely on one translator, indicating that diversifying training data, even across translators, contributes positively to model robustness.

For instance, combining Ibn al-Jazzar and Ibn at-Tilmīdh for training shows an F1-score of 83.58 on the Ibn at-Tilmīdh test set, indicating an improvement as compared to training solely on Ibn at-Tilmīdh (82.34), or on manuscripts from Oliver Kahl only, namely the Ibn at-Tilmīdh and Sabur ibn Sahl (81.70). The positive effect is also seen when we equally mix all manuscripts, achieving 83.60 F1-score on the best test (Ibn at-Tilmīdh).

These scores indicate a noticeable improvement in performance which suggests that training on diverse data improves the model's ability to generalize across different texts.

Effect of Training Dataset Size

The effect of training dataset size was evaluated by comparing the results from the equal training dataset size experiments with those from the full manuscripts (Table 6) as opposed to the previously discussed experiments with equal training.

Observing the F1-scores, a pattern becomes noticeable: as more data is used for training through employing all manuscripts, the model's performance shows an improvement across different test manuscripts. However, a closer examination reveals that while increasing the training data positively impacts results, there's a relative slow-down in the rate of performance gains as the dataset enlarges. This suggests that while enlarging the training pool is beneficial, there is a tipping point beyond which additional data does not boost the NER model's performance so much. This reflects a diminishing return on investment when consistently escalating the amount of training data. For instance, training on all manuscripts with equal sizes resulted in an F1-score of 75.91 for Ibn al-Jazzar, 83.60 for Ibn at-Tilmīdh, and 83.21 for Sabur ibn Sahl. In contrast, using the full manuscripts yielded higher scores: 80.91 for Ibn al-Jazzar, 86.51 for Ibn at-Tilmīdh, and 87.21 for Sabur ibn Sahl. This trend illustrates the benefits of larger training datasets while indicating that beyond a certain point, the gains in performance decrease.

3.4 Error Analysis

The error analysis aims to study the performance of the DeBERTaV3 model on the Named Entity Recognition task, focusing on specific tags and instances where the model's predictions deviated from the expected annotations. This section examines the precision, recall, and F1-scores for each tag, from the results of t the best performing model that was trained on all manuscripts, and analyzes the performance through detailed examples from remedies from Ibn

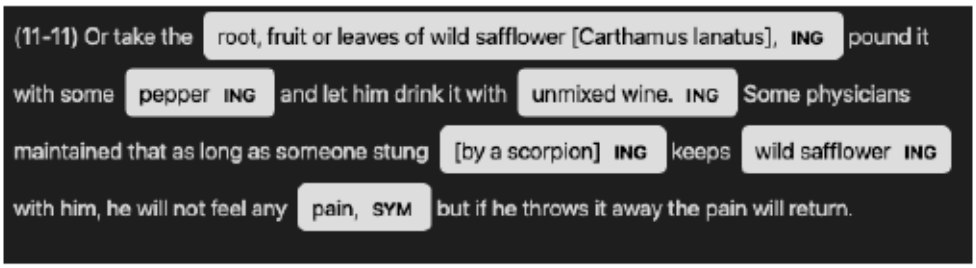
al-Jazzar's pharmacopeia, which had the lowest F1-score among the manuscripts across the different discussed experiments.

Table 7 shows that Ingredient entities have the best score (0.92), and highest support (1675). Organ entities have a much lower performance compared to the other tags (0.69), and this is possibly due to being the least represented entity in the support (62). Additionally, Type entities (0.85) perform better than Symptom entities (0.80) even though they have lower support (249 and 335, respectively). This may be due to the tendency of Types entities to be simpler in form as single-word entities such as "pasille" and "pill", while symptoms can in many instances be multi-word entities such as "flaming sensations", "remnants of fevers" and "urinating blood and purulent matter".

Tag	Precision	Recall	F1-Score	Support
ING	0.91	0.92	0.92	1675
ORG	0.64	0.74	0.69	62
SYM	0.77	0.83	0.80	335
TYPE	0.87	0.84	0.85	249

Table 7 - Tags-specific performance

To further understand the model's performance, we analyze three specific examples from Ibn al-Jazzar's pharmacopeia which shows both correct and incorrect predictions:



Example 1

In the first example, the model correctly identifies multiple ING tags such as “root, fruit or leaves of wild safflower [Carthamus lanatus]” “pepper” “unmixed wine” and “wild safflower”.

However, it incorrectly identifies "by a scorpion" as an ING instead of recognizing it as part of the symptom or context. Additionally, "pain" is correctly tagged as SYM, but the sentence structure causes some confusion, leading to potential misclassification of related terms.

(17-2) Galen said that he has not seen anyone recover from this disease [i.e., leprosy] SYM when it has reached its climax, except for the leper who drank wine ING in which there was a viper that had been chopped into pieces. He was stripped of his [diseased] skin because he drank from that wine. ING Because of this eyewitness I know that the statement of the former [physicians] that the fat and flesh of vipers ING is beneficial against this disease and against all diseases that are hard to heal is a true one.

Example 2

In the second example the SYM tag for "leprosy" is correctly identified, but "skin" was not identified with the ORG tag, which might be due to its low score caused by its less frequent nature in the original annotations.

(26-7) A powder TYPE beneficial for a head wound SYM and for someone struck by a sword, stabbed by a knife or beaten by a stick: Take one ounce each of aloe, ING myrrh, ING gum ammoniac ING and sarcocol, ING two mithqals each of verdigris ING and ?abarzad sugar. ING Pound these ingredients well, then fill up the wound SYM or the head wound SYM with it, moisten a piece of fine cloth with saliva ING and put it on [the wound]. SYM Sprinkle some of the powder TYPE on it and bandage TYPE it with a rag until it is healed.

Example 3

In the third example, the model performs well in tagging ING entities like "aloe" "myrrh" "gum ammoniac" "sarcocol" "verdigris". It also correctly tags "powder" as TYPE and "head wound" as SYM. However, the term "bandage" is tagged incorrectly as a TYPE while it is a verb of action, which we do not aim in extracting in this task. "abarzad sugar" was not fully considered as an ING, only "sugar" was tagged, which indicates that entity boundary issues may occur.

Several common error patterns emerged from the analysis. One issue was contextual ambiguity. The model tends to struggle with context-specific interpretations, especially when the same term appeared in different contexts. Entity boundary errors were also present. The model occasionally misidentified the boundaries of entities such as for multi-word terms.

The model also struggled with rare and domain-specific terms. Terms specific to historical medical texts, such as archaic names for ingredients or rare symptoms, posed some challenges.

The ORG tag, with the lowest F1-score, exemplifies this issue, as the model struggled with less common references to body parts.

Lastly, complex sentence structures typical of pharmacopeias contributed to errors in entity recognition. Some lengthy sentence constructions made it challenging for the model to go through long, complex sentences with multiple clauses and embedded contexts.

3.5 Discussion

Regarding what was seen in the error analysis, to address the encountered issues, improved preprocessing steps could be implemented to normalize archaic terms and ensure consistent representation of entities across different contexts. Expanding the annotated training dataset to include more examples of rare and domain-specific terms would provide the model with better exposure to the types of entities it needs to recognize. Developing post-processing rules or heuristics to correct common errors, such as entity boundary mistakes and misclassifications like for repetitive terms and complex sentence structures would also help in refining the model's output.

This error analysis reveals that while the DeBERTaV3 model performs well overall, there are specific areas where its performance can be improved. Addressing contextual ambiguity, improving the training dataset through diversification and balancing tag counts, and implementing better preprocessing and post-processing strategies would be the necessary steps for improving this NER task overall.

Further, the analysis of the training results reveals several insights into the performance and generalizability of transformer-based NER models for historical medical texts. Regarding the transferability between translators, the variability in performance across different translators underscores the importance of considering translation styles in NER tasks. While the model can generalize, the drop in performance indicates the need for a model that can better handle stylistic differences, this may be achieved through a wider diversification of the training data. Indeed, the differences in entity distribution and writing styles between translators/authors affect the model's performance which suggests that diverse authorship can achieve better generalization.

Combining multiple manuscripts for training does improve the model's robustness and exposure to diverse linguistic and contextual patterns helps the model adapt to new texts more effectively. Despite this fact, and while larger training datasets improve model performance, it

seems the benefits taper off beyond a certain point. This finding shows the importance of balancing data quantity and quality in the NER task.

These insights inform future research directions and shows the need for diverse training datasets to fine-tune a more robust NER models for historical texts. Additionally, this study demonstrates the potential of transformer-based models to facilitate the analysis of historical medical manuscripts through automated recognition of named entities that are required for the wider goal of insight extraction from these old pharmacopeias.

While the results of this study demonstrate the potential several limitations and challenges were encountered throughout the research process, such as annotated data scarcity. Historical pharmacopeias are relatively not abundant, and annotating these texts requires a big amount of time (around 1 month of part-time work per manuscript) and domain expertise. Likewise, the differences in authorship, translation and editing styles pose a challenge in that each unique approach to rendering historical texts into English introduced variability that the models had to account for. This variability often led to inconsistencies and affected the model's performance.

The archaic language and specific medical terminology used in historical texts added another layer of complexity. The models needed to be able to recognize terms that are less commonly used in general-purpose English-language transformer models such as DeBERTaV3, which required careful preprocessing and annotation. Indeed, ensuring high-quality and consistent annotations was very important. The process of manually annotating texts is time-consuming and prone to human error, despite the efforts to maintain consistency through the expertise of the annotation team.

Regarding Transformer-based models, while reliable, they are computationally expensive. Training these models requires intensive computational resources, which can be a barrier for many researchers.

3.6 Conclusion

This chapter studied the application of transformer-based models for Named Entity Recognition on historical medical texts, focusing on manuscripts translated to English. The research aimed to identify the best-performing model and assess its generalizability across different manuscripts and translation styles. The first phase of the study compared several transformer-based models, including BERT, RoBERTa, XLM-R, BioBERT, and DeBERTaV3. This study showed that DeBERTaV3 emerged as the best-performing model,

achieving the highest F1-score of 86.03. Subsequent experiments evaluated the model's performance in various training setups, including single manuscript training, combining two manuscripts, and combining all manuscripts to test the generalizability of the NER task.

For these experiments, the results demonstrated that while the model could generalize across different texts, its performance was influenced by the translation and editing style of the manuscripts. Combining multiple manuscripts for training improved the model's robustness and adaptability through the diversification of the training dataset. Additionally, increasing the training dataset size improved performance, although with diminishing benefits proportionally to size increase.

Chapter 4: Linking Vernacular Plant Names to Their Taxa

4.1 Introduction

While the previous task, Named Entity Recognition, identifies and categorizes entities in texts, it is the process of Named Entity Disambiguation and Linking that assigns these entities to specific, unambiguous identifiers of data in a data infrastructure which provides more meaning to the textual data. This is why the NED and NEL tasks are of high significance in our Text Mining pipeline. In texts where the same entity might be referred to in multiple ways or where different entities might share similar names, this task is important in order to accurately represent information. For example, different instances of plant names across centuries, geographic locations and languages necessitate precise disambiguation to ensure accurate linkage to modern scientific data infrastructures.

Indeed, without effective disambiguation and linking, the potential insights these texts offer about historical medical practices and their relevance to modern medicine could remain obscured. Furthermore, the correct linking of entities to contemporary botanical resources allows researchers to study their modern-day implications by accurately representing the information present in a pharmacopeia in a way that would allow querying to investigate the most common combinations of plants in a text.

This chapter aims to achieve several objectives within the broader aim of analyzing old pharmacopeias. The primary objective is to develop and assess, through using existent tools, a pipeline of disambiguation and linking for entities identified by the NER process. Given the ambiguity in natural language text, developing such a pipeline proves useful. This integration is necessary for creating a unified system that can process our corpora from raw input to linked, structured data. Throughout the chapter, our methods and the tools that we use will be described and critically evaluated.

Finally, this chapter aims to lay the groundwork of completing the dataset that will be stored in a graph database (5.2.2) in the aim of completing the pipeline of pharmacopeia mining and analysis.

4.2 Resources and Data Preprocessing

Here, we discuss the resources and methodologies employed to process the data for the NED and NEL tasks. We describe the tools that we utilized, the specific preprocessing steps taken to clean and align the data, and the overall approach to ensuring that the entities recognized by the NER process are accurately identified and ready for analysis.

The choice of reference resources is important for the feasibility of the NED/NEL process. The following are the main resources used in this research:

BabelNet and Babelfy

As previously discussed in 2.2.4, BabelNet [83] is a multilingual encyclopedic dictionary and semantic network that integrates information from various sources, including WordNet, Wikipedia. It is especially useful in this research for handling the semantic disambiguation of entities. BabelNet provides a wide set of concepts and relationships which makes it possible to match historical terms with their modern equivalents or related concepts in a wide array of languages. This in turn makes it a suitable intermediary tool in our goal of linking a plant mention to its scientific taxon in a different resource, since it is linked to many different intermediaries such as Wikidata which in its turn is linked to specialized data infrastructures.

Babelfy is a tool built on top of BabelNet that provides an API for word sense disambiguation. It processes text input to identify and link entities in-context to BabelNet concepts, which offers an efficient way to disambiguate terms and connect them to the entities present in BabelNet that we discussed. Indeed, this tool is used to initially process the text extracted from the historical pharmacopeias, identifying potential BabelNet IDs for each recognized entity.

Wikidata

Wikidata [118] is an open data infrastructure that acts as a central storage for structured data across various Wikimedia projects. It offers an exhaustive repository of information about people, places, concepts, even plants which makes it a suitable resource for linking our vernacular plant mentions as identified with the BabelNet ID to their standardized counterparts. In the context of the pipeline, Wikidata serves as an intermediary between the historical and vernacular entities identified by BabelNet from the pharmacopeias and modern scientific resources like the Global Biodiversity Information Facility.

Global Biodiversity Information Facility (GBIF)

Global Biodiversity Information Facility (GBIF)[14] is an open-data infrastructure that gives access to data about all forms of life on Earth. GBIF operates as a network of participating countries and organizations that share biodiversity data openly and freely. It is designed to serve as a global repository for biodiversity data and encompasses a large array of species, from plants and animals to fungi and microorganisms.

The primary function of GBIF is to provide a single point of access to millions of biodiversity records, which are contributed by various data publishers, including museums, herbaria, universities, government agencies, and non-governmental organizations. These records typically include information on species' taxonomic classification, geographic distribution, and specimen details, among others. GBIF's dataset is one of the largest and most exhaustive collections of biodiversity data, comprising nearly 2 billion occurrence records that can be used for a wide range of scientific purposes. According to its website, around 10,000 peer-reviewed papers have used its datasets which makes it a widely accepted resource.

In the context of plants, GBIF maintains a taxonomic backbone that standardizes species names across different sources. One of its notable features is the detailed geographic distribution data, which shows the coordinates of where species have been recorded globally. This data is useful for disambiguation because it allows to cross-reference the historical context of a plant's mention (e.g., a text from Baghdad or Kairouan in our case) with its known geographic distribution. GBIF was integrated in the pipeline as it is a suitable tool for reaching the final link to the scientific name of a plant mention.

Data Preprocessing

As established in previous chapters, historical texts often contain noisy, inconsistent, and context-dependent language. Taking this into account is also relevant in the case of this pipeline. The preprocessing phase aims to clean the data and transform it into a format that is more suitable for accurate disambiguation and linking.

The first step in preprocessing, is in common with the one performed for the first NER task, which involved cleaning the text extracted from the pharmacopeias by removing unnecessary formatting, punctuation, and other non-alphanumeric characters that could interfere with the

[14] Openly accessible via https://www.gbif.org/

entity recognition and linking processes. Tokenization was then performed to break the text into individual words or phrases to make it easier to identify entities as was done in the NER task.

One of the challenges in working with old texts is the variability in how entities are mentioned. For example, an ingredient might be referred to with additional descriptive words, such as "grilled apple seeds" where only the word "apple" is necessary for accurate linking as a taxon in modern data infrastructure. Indeed, old pharmacopeias frequently describe plants in terms of their parts (e.g., seeds, leaves) or transformations (e.g., grilled, dried), it was necessary to remove these qualifiers to focus on the primary entity. To do this, two specialized hand-crafted dictionaries were created, based on the manual assessment of the ingredient entities in the pharmacopeia of Sabur Ibn Sahl: one focused on identifying and removing parts of plants and another on removing transformation-related keywords. These dictionaries were designed to strip away non-essential words from entity mentions, allowing the linking process to focus only on the core term and which improves the likelihood of successful matches in BabelNet.

The parts dictionary includes terms such as "leaf", "root", "seed", and "flower" among others, which refer to specific parts of plants. For example, if a pharmacopeia referred to "cucumber seeds" the term "seeds" would be removed, leaving "cucumber" as the entity to be linked. The transformation keywords dictionary contains terms related to the processing or transformation of plant materials, such as "dried", "boiled", "grilled", and "peeled". Like the parts dictionary, this dictionary was used to remove these transformation-related keywords from the entity mentions. For instance, "dried apple" would be simplified to "apple" before the linking process, ensuring that only the essential term is used in the NEL pipeline. Figure 21 illustrates an example this process using the previously mention example of "cucumber seeds".

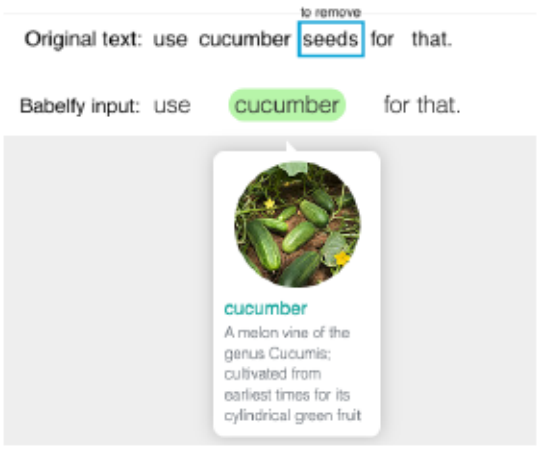


Figure 21 - Example of preprocessing before using Babelfy

The sections 5.3.1 and 5.3.2 go into more detail regarding the words that were used in the dictionaries, as this task was also useful in the context of information modeling because information about the parts of plants and transformations were included in the model and have a direct relationship with the essential term that we treat in this section regarding disambiguation.

This step ensured, by reducing noise, that Babelfy worked only with the essential term to improve the chances of finding a correct match in BabelNet and other resources.

4.2.1 The Pipeline

Following preprocessing, the Babelfy tool is employed on the plain text for the initial disambiguation. Babelfy uses the semantic network of BabelNet by using the context provided by surrounding text to accurately disambiguate entities. This is useful in cases where an entity could have multiple meanings, such as "apple" which might refer to both a fruit and a company. Babelfy's ability to consider the broader textual context allows it to select the most relevant meaning, linking the entity to the appropriate BabelNet concept. Although the pipeline was built upon the APIs of the different resources and coded using Python, we use the web interfaces of the different tools as visual support in the following. Figure 22 showcases the first step where Babelfy identifies entities based on the plain text of the corpus. In this case, we use the example of the "alhagi" plant going through the pipeline.

Figure 22 - First step of the pipeline : word sense disambiguation using Babelfy

In this process, we focus on the disambiguation results of entities that are tagged as ING (ingredients) from the NER, because Babelfy disambiguates all entities that it finds in BabelNet, including words that we do not use in our analysis. After finding the BabelNet ID of the entities through Babelfy's disambiguation tool, the next step of the process involved mapping the disambiguated entities to their corresponding entries in Wikidata which would be the intermediary.

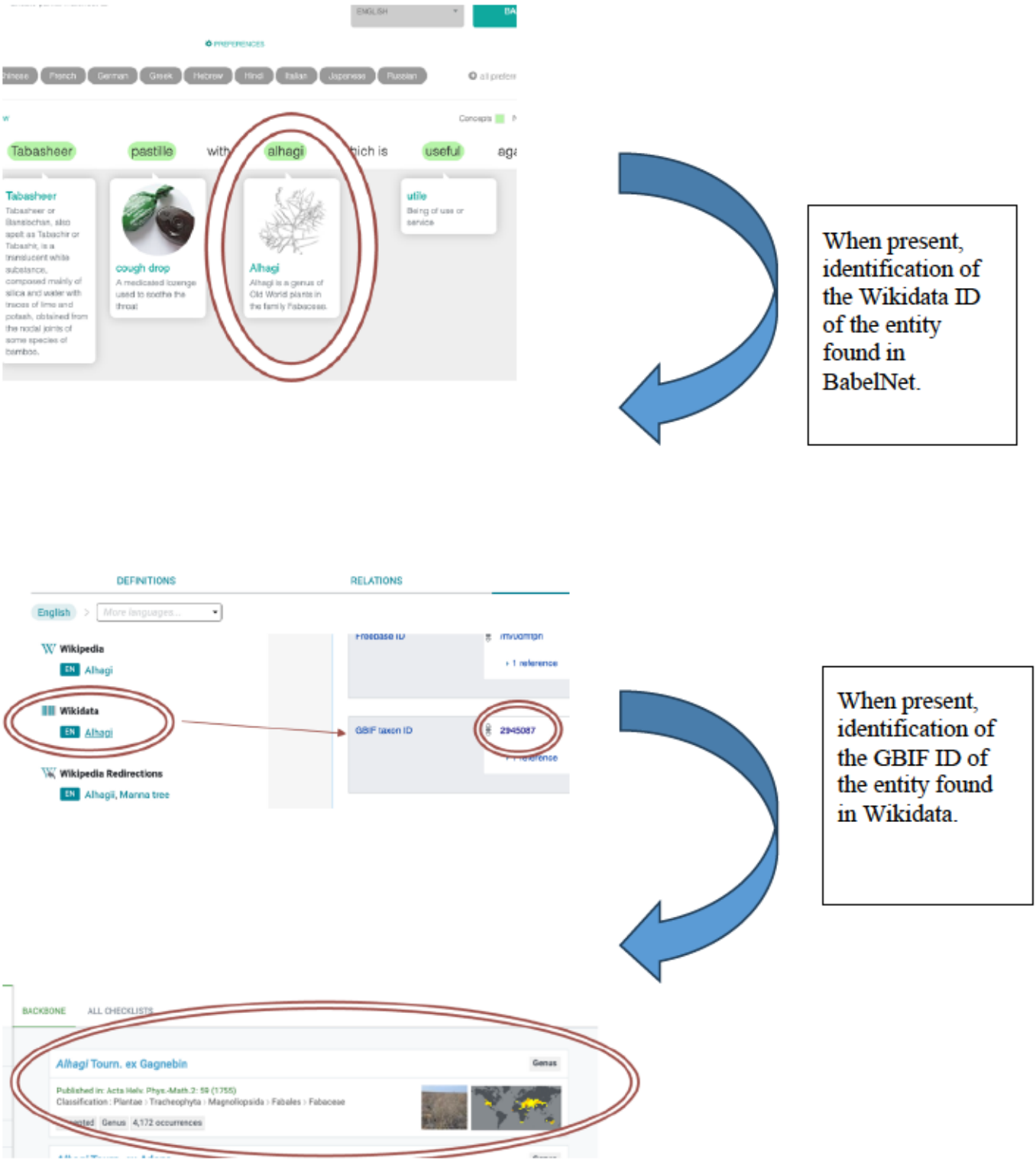


Figure 23 - Next steps: finding Wikidata and GBIF IDs

While semantic disambiguation through tools like Babelfy are effective to produce candidate links, it is not always sufficient when dealing with plant names where the same term might refer to different species based on geographical context. The following figure shows the page of the "alhagi" plant from the GBIF that was found on Wikidata as a result of the process

through the pipeline from Babelfy. A total of 6 different species (children taxa) of Alhagi are mentioned by GBIF, 3 of which have their own geo-reference coordinates and 3 others not as well-documented.

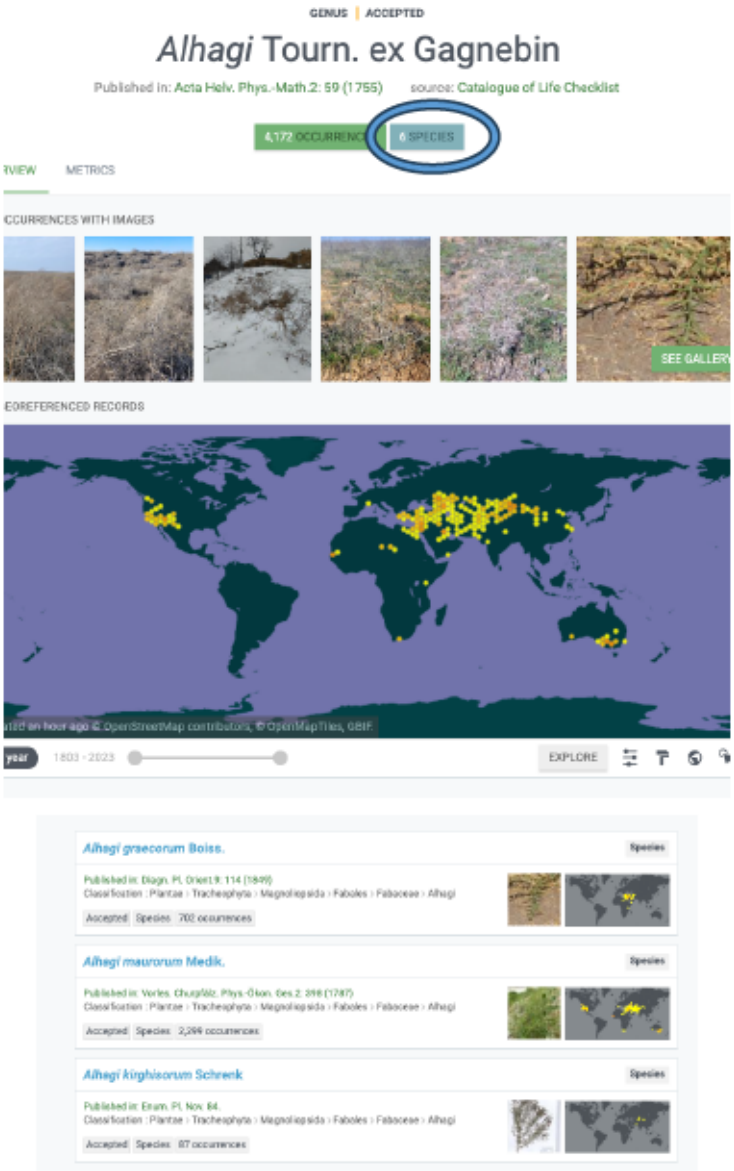


Figure 24 - Geo-references and candidate species of Alhagi

Indeed, a plant's vernacular name can be linked to different species depending on the geographic context. This is where the novelty of the geographic filtering step is used. Old pharmacopeias often reference plants that are commonly found in the region where the text was written. For example, a manuscript originating from Baghdad is more likely to reference plants that are native to the surrounding regions of Mesopotamia rather than Africa or the Americas. To address this geographic specificity, a custom geographic filtering algorithm and scoring system was developed. The algorithm utilizes geographic distribution data from GBIF

to prioritize and disambiguate plant entities based on their likelihood of being found in the region associated with the manuscript. For instance, if a plant entity identified in the text has multiple possible children taxa in GBIF, the algorithm will prioritize those matches that are geographically relevant to the manuscript's origin[15].

The geographic filtering process works by first identifying the geographic coordinates or region associated with the manuscript's origin. This information is then used to filter potential matches in GBIF, assigning higher scores to entities that have a higher density of documentation in that region. For example, in the case of a Baghdad-originating manuscript, the algorithm would give preference to plant species that are commonly found a Middle Eastern radius around Baghdad. This approach augments the granularity of the disambiguation process and alleviates the factor of human choice error if this task were to be performed manually. Indeed, making an educated choice when selecting each of the species candidates would be time consuming, especially when some vernacular plant names have hundreds of identified species names. The scoring system was designed to make use of the geographic coordinates of references contained within GBIF. The core of this system is a weighted scoring mechanism that prioritizes taxa based on their occurrence density within specified distances from a reference location. The advantage of this system is its applicability no matter where the analyzed manuscript originates from, as only the central coordinates need to be changed in accordance with a manuscript's origin. For instance, in order to calculate the distance of a coordinate from Baghdad (having the coordinates 33.3, 44.4) we use the Haversine Distance formula as expressed in [119].

This formula is utilized to compute the great-circle distance *d* between two points on the Earth's surface. The formula is given by:

$$d = 2R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

where:

R is the Earth's radius (approximately 6371 km) [119],

ϕ_1 and ϕ_2 are the latitudes of the two points in radians,

[15] Python code accessible via https://github.com/karimhaff/pharmacopeias/blob/main/best_taxon

$\Delta\phi$ is the difference between the latitudes,

$\Delta\lambda$ is the difference between the longitudes of the two points.

In our case, this distance metric is applied to calculate how far each species occurrence is from Baghdad.

Radii Selection

We chose a scoring system that evaluates species occurrences within three distinct radii: 50 km, 500 km, and 1500 km. These radii are used to assess the density of occurrences at different scales of proximity, ranging from local to regional distributions.

- 50 km radius: Focuses on species that are extremely localized around Baghdad.
- 500 km radius: Captures species with a broader but still regionally significant presence.
- 1500 km radius: Encompasses species distributed over a much larger area, covering surrounding regions.

Density Calculation

For each taxon t, the density ρ_{tr} within a radius r is calculated using the formula:

$$\rho_{t,r} = \frac{n_{t,r}}{\pi \cdot r^2}$$

Where:

- r is the radius under consideration

- n_{tr} is the number of occurrences of taxon t within radius r.

Normalization and Coefficient Application

The calculated density for each taxon within a specific radius is then normalized by comparing it to the maximum density observed for any taxon within that radius. This normalized value is multiplied by a predefined coefficient C_r associated with the radius:

$$S_{t,r} = \left(\frac{\rho_{t,r}}{\rho_{max,r}}\right) \cdot C_r$$

Where:

- ρ_{tr} is the density of taxon t within radius r,

- ρ_{maxr} is the maximum density observed for any taxon within radius r to ensure comparability across different radii by scaling against the maximum observed.,

- C_r is the coefficient assigned to that radius (from most to least likely: 0.9 for 50 km, 0.6 for 500 km, 0.3 for 1500 km).

Selection of Maximum Score

The overall preliminary score for each taxon t is determined by selecting the maximum score across all radii:

$$S_t = \max_r S_{t,r}$$

Essentially, it ensures that the taxon's final score is the highest score it achieved within any of the radii.

Adjustment for Total Occurrences

To account for the overall occurrences of the taxon across all radii, the score is adjusted by considering the proportion of occurrences relative to the maximum observed across all taxa:

$$O_t = \frac{n_t}{n_{max}} \cdot 0.1$$

Where:

- n_t is the total number of occurrences of taxon t across all radii,

- n_{max} is the maximum total occurrences observed among all candidate taxa.

The purpose of this adjustment is to give a 10% boost to taxa that have a higher overall number of occurrences, ensuring that taxa which are generally more prevalent are slightly favored, though not overwhelmingly so.

Final Score Calculation

The final score S_{ot} for each taxon is obtained by adding the total occurrences adjustment O_t to the maximum preliminary score S_t:

$$S_{final} = S_t + O_t$$

Overall, this scoring system ensures that taxa with the highest density within the smallest radius (50 km) receive the most boost in their scores, reflecting their likely relevance based on geographic proximity. The adjustment for total occurrences across all radii helps refine the ranking, though the primary influence remains the density within the smallest radius.

In the case of the Alhagi plant that we have been using as an example, the species of *Alhagi maurorum* (GBIF ID: 2945092) received the highest score of 0.7. We compare the results of this taxon with two other candidates (*Alhagi graecorum* and *Alhagi pseudalhagi*) for each equation in Table 8 adjusted to 4 decimal places, except for the density ρ_{tr} which is a small value.

Taxon ID	Radius (km)	ρ_{tr}	S_{tr}	S_t	O_t	S_{ot}
2945092 - Alhagi maurorum	50	0.0	0.0000	0.6000	0.1000	**0.7000**
	500	5.092958178940651e-06	0.6000			
	1500	4.031925224994682e-05	0.3000			
2945091 - Alhagi graecorum	50	0.0	0.0000	0.6000	0.0958	**0.6958**
	500	5.092958178940651e-06	0.6000			
	1500	3.8621599523633275e-05	0.2874			
11374752 - Alhagi pseudalhagi	50	0.0	0.0000	0.2063	0.0678	**0.2741**
	500	0.0	0.0000			
	1500	2.7728327863121322e-05	0.2063			

Table 8 - Comparison of 3 taxa candidates for alhagi

Additionally, Figure 25 illustrates the geographic distribution of its coordinates that were generated on a map using the Folium library on Python[16].

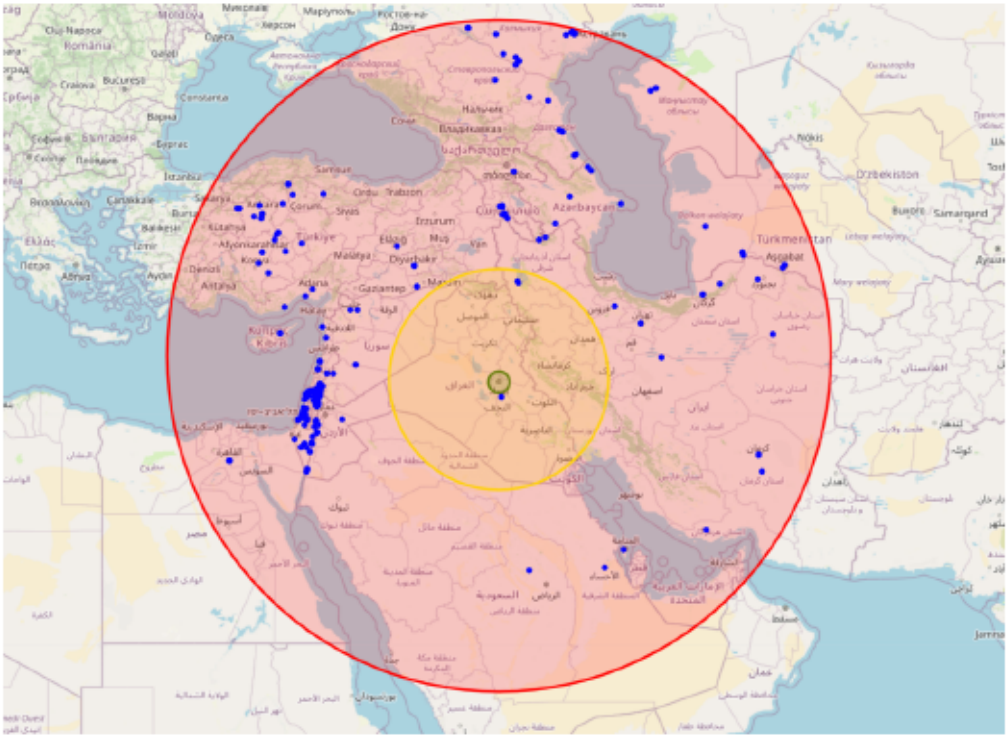


Figure 25 - Alhagi mauromum's geographic distribution based on GBIF coordinates.

In this figure, the green circle represents the 50 km radius, the yellow circle represents the 500 km radius, and the red circle represents the 1500 km radius. We notice that the best scoring candidate does not have any coordinates within the 50 km radius but does so in the 500 km one, which indicates a minimum score of 0.6 that was improved by 0.1 as it would have the highest relative density, as calculated by the applied formula. This step is applied to all the GBIF IDs when present in order to map the best candidates to the original plant name that began its processing with Babelfy at the beginning of the pipeline. Figure 26 and Figure 27 illustrate the geographic distribution of the second and third ranked candidates which have lower scores.

[16] https://pypi.org/project/folium/

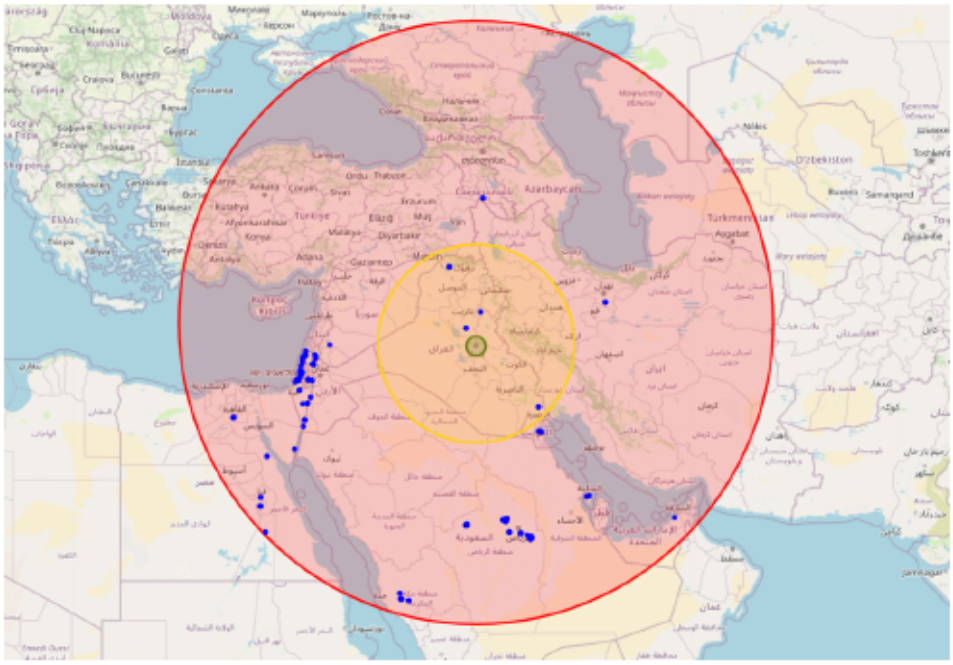


Figure 26 - Alhagi graecorum's geographic distribution based on GBIF coordinates.

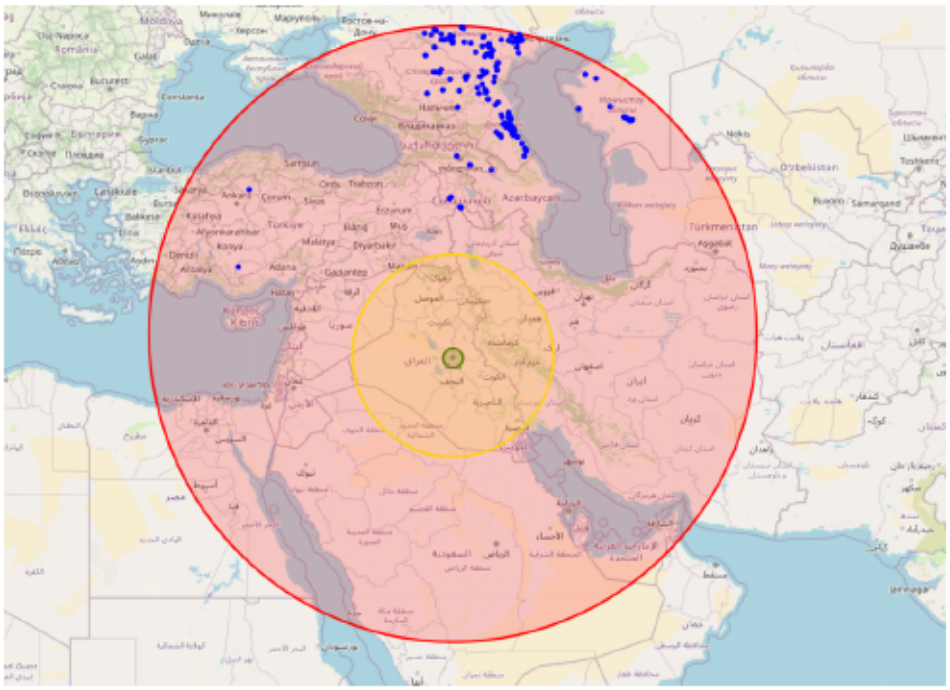


Figure 27- Alhagi pseudalhagi's geographic distribution based on GBIF coordinates.

4.2.2 Performance Analysis

The processes we developed were applied to first set of 1,000 ingredient entities that appear in Sabur Ibn Sahl's pharmacopeia. Indeed, the corpus was cut in order to form a sample that

contains the first 1000 entities that were tagged as ING in the NER task. Out of the 1000 entities, we note that 820 of them were plant-based.

When the text was processed through Babelfy for disambiguation, out of the 820 entries, 767 were analyzed and 53 were not detected (notably Arabic transliterations such as "zanbaq, diyāqūḏ, kaukab" or rarer appellations like "gum-arabic"). This high coverage of 93.54% shows the robustness of BabelNet in detecting a diverse range of terms as named entities. However, the process was not without errors. Notably, the disambiguation phase encountered issues with homonyms such as words like "rose" being misinterpreted as the verb "to rise," leading to incorrect or failed links in subsequent stages. Indeed, we count 60 of such cases (7.32%) where Babelfy incorrectly identified entities as being out-of-domain verbs or adjectives.

Following BabelNet, 572 of the 820 entities were successfully linked to Wikidata, representing a notable drop-off. This reduction by approximately 25.42% underscores the limitations in the availability of Wikidata identifiers for many entities disambiguated by BabelNet. This stage emerged as a bottleneck, where even if BabelNet could process an entity, the lack of a corresponding Wikidata identifier not present in BabelNet prevents further linking.

The final linking stage to GBIF revealed further attrition, with 266 of the 572 Wikidata-linked entities successfully connected to GBIF. This represents an additional reduction of 53.50%, reflecting the challenges of linking vernacular botanical names to modern taxonomies and the limitations of GBIF's coverage, as Wikidata does not always contain the GBIF ID for all plants. Despite this, the use of a geographic filtering algorithm during this stage proved useful, as it prioritized matches based on the geographic origin of the texts. For instance, species referenced in the Baghdad region were correctly identified (relative to the confidence towards data availability on GBIF) whenever all the necessary identifiers were present through the different steps of the pipeline, which validates the algorithm's effectiveness in narrowing down potential candidates, while keeping in mind a degree of uncertainty.

4.2.3 Discussion

The NED/NEL pipeline, while effective in many respects, faced obstacles that impacted the overall success of entity linking. These challenges can be categorized into three primary areas: error propagation from the disambiguation process, data attrition at each stage of the pipeline, and the limitations of the used resources.

Error Propagation from Disambiguation

One of the most notable issues encountered in the pipeline was the propagation of errors originating from the Babelfy and BabelNet disambiguation stage. As noted, Babelfy was able to successfully analyze 767 out of 820 plant-based entities. However, the case of "rose," misinterpreted as the verb "to rise", or that of "camphor" that was detected as the adjective "camphorated", are notable examples of how such errors can have a cascading effect. Once an entity was incorrectly disambiguated, the subsequent stages, linking to Wikidata and GBIF, were to fail. This problem was present for homonyms and words with multiple meanings, where Babelfy's context-insensitive approach struggled to distinguish between different senses of a word.

The cascading effect of such errors meant that inaccuracies at the disambiguation stage caused successive losses which led to a higher rate of entity loss as the pipeline progressed. Indeed, this sheds light onto a limitation in the current approach: the reliance on a single disambiguation resource (BabelNet) that, while exhaustive, is not without its flaws. The pipeline would benefit from incorporating additional contextual analysis or even alternative disambiguation resources that could cross-verify the interpretations provided by Babelfy.

Data Attrition Across Stages

Another challenge was the attrition of entities as they moved through the pipeline. Starting with 820 entities, the process saw a reduction to 767 at BabelNet, 572 at Wikidata, and finally 266 at GBIF, showing an overall success rate of 32.44%. This steady decline illustrates the difficulties in maintaining a high level of entity retention across multiple stages of processing. The subsequent drop-off between BabelNet and Wikidata, where 195 entities were lost, suggests that while Babelfy was able to disambiguate a term, it did not always correspond to an entry in Wikidata. This points to a limitation in Wikidata's coverage and a lack of links between BabelNet entries and their Wikidata equivalent.

The most attrition occurred in the final stage, where only 266 entities were successfully linked to GBIF. This stage was most probably affected by the lack of links from within Wikidata to the GBIF identifiers that would make possible the geographic analysis. The system's inability to retain a larger proportion of entities through to the final stage suggests that further improvements are needed in both the disambiguation process and the contents of the used data infrastructures.

Resource Limitations

The reliance on modern resources like Wikidata and GBIF introduced its own set of challenges. These resources are primarily designed to handle contemporary data and may not fully encompass the historical niche or vernacular language required for our research. The fact that several entities could not be linked to Wikidata or GBIF shows the limitations in their coverage. The fact that a long pipeline requiring many resources was needed to perform this task shows the specialized aspect of the task, and demonstrates that, while the resources we used had non-negligeable success in disambiguating and linking plant entries, they were not designed with our pipeline in mind.

Indeed, the lack of corresponding entries in Wikidata or the absence of their relevant taxonomic GBIF identifiers meant that even well-disambiguated entities could not be effectively linked. This limitation shows the need for more specialized resources or the expansion of the existing resources to improve the accuracy of our pipeline and reduce the shortcomings of a complex pluralistic process.

Indeed, one of the most notable needs identified in this chapter is the development of an exhaustive resource for vernacular plant entities. Such a resource could be created in collaboration with experts in history and botany to ensure that it finds the full range of entities mentioned in historical texts and aligns them with modern scientific classifications. However, the robustness of the multilingual aspect of BabelNet must not be discarded, because while we focus on English translations of Abbasid pharmacopeias, the methodologies developed here could be applied to texts in different languages, such as Latin, Greek, Chinese or in the original Arabic of the medical manuscripts, as BabelNet has the potential to reach a link to botanical entries in data infrastructures no matter the language of origin of the pharmacopeia. Exploring how these techniques perform across different languages would provide clearer insights into the generalizability of the NED/NEL system that was developed, to check if a language-agnostic pipeline is possible.

Additionally, the success of the geographic filtering approach suggests that further refinements could yield even better results. Future work could explore more complex algorithms for geographic filtering, such as those that take into account historical trade routes related to the manuscript's region of origin, climate zones, or ecological niches, to improve the accuracy of plant entity linking even further by optimizing the selection of candidates through additional

boosts to the scores based on such knowledge that can be acquired through collaborations with historians and botanists.

Indeed, the techniques developed here could be adapted for use in other domains where historical text analysis is necessary, such as the study of ancient trade records. The integration of NED/NEL systems with geographic filtering and other contextual tools could have different use cases that may be useful for analyses in these areas.

4.2.4 Conclusion

The development of the NED/NEL pipeline for processing old pharmacopeias represents a step forward in the overall aim of analyzing such texts, through the larger pipleline of Text Mining. The integration of disambiguation methods of existing resources such as BabelNet with a novel approach like geographic filtering has demonstrated the potential to accurately link plant-based entities to modern scientific resources, despite the challenges posed by errors, data attrition, and resource limitations.

All in all, considering the system's current iteration has shown promising results, it also showed the need for further refinement and expansion, through improving disambiguation techniques, expanding the contents of resources, and incorporating additional contextual layers into the NED/NEL pipleline. Building on this pipeline and improving its accuracy would allow for the automation of the insight extraction from old pharmacopeias and would be a necessary complement to the preceding NER task and base upon which will be built the subsequent Information Modeling task.

Chapter 5: Information Modeling for Old Pharmacopeias

5.1 Introduction

Old pharmacopeias contain complex relationships between their entities. As presented in the previous steps of the Text Mining pipeline, the terminology within these texts is often inconsistent as the text is written in natural language. Addressing such difficulties to reach queryable data requires a structured approach to information modeling. Indeed, information modeling is necessary in the context of a systematic analysis of the old pharmacopeias that we work on. In such texts, information modeling would involve organizing data to accurately reflect the different relationships between instances (originating from named entities) within the texts. The variability and complexity in historical texts, where there is a need to group vernacular terms into standardized forms, shows the need for a well-designed data model in order to manage these complexities and to support the subsequent analysis.

As was reviewed and concluded through the literature, using graph databases would be a suitable approach to attain such queryable data from unstructured texts. Unlike traditional relational databases, graph databases provide efficient data-traversal when the relationships are complex. By representing named entities such as remedies, ingredients, organs, and symptoms as database instance nodes, and their relationships as edges, graph databases would allow for a adequate representation of the data. The graph model aids in the identification of patterns and connections within the data, supporting the analysis of the underlying medical practices and knowledge systems in historical contexts.

The objective of this chapter is to detail the development of a graph database model designed specifically for old pharmacopeias. The chapter will outline the design decisions underlying the database, with a focus on addressing the encountered difficulties.

The chapter will begin by discussing the rationale for selecting Neo4j as the graph database technology and will then provide an overview of the core instances and relationships within the database, based on the designed model. We then proceed to address the challenges encountered during the modeling process, including the representation of ingredient transformations and the integration of plant parts as distinct database instances. Afterwards, the potential methods of insights extraction through Formal Concept Analysis will be described, as this task constitutes the final step of the Text Mining pipeline.

5.2 Design of Graph Database Model

5.2.1 Selection of Graph Database Technology

In the development of a graph database model, while the selection of the appropriate database technology is secondary to the main task of modeling, it remains important as it is the medium where the data is to be practically stored. Here, we present the resource Neo4j, an open-source graph database implemented in Java, that was chosen for this project due to its ease of use for data non-specialists, as well as its capability to manage and store data in the form of graphs rather than tables.

Neo4j is described as a fully transactional database and a persistent Java engine, designed to store and manage graph structures efficiently [120][17]. Unlike relational databases, Neo4j utilizes native graph storage which makes it suitable for applications that deal with large datasets. The selection of Neo4j was further influenced by its user-friendly interface (Figure 28), which facilitates accessibility for non-specialists who are the main benefactors of the Text Mining pipeline, such as historians and biologists, who may wish to study the data without extensive technical expertise. Neo4j's interactive user experience allows users to intuitively traverse through complex datasets, visualize relationships, and execute queries with minimal technical knowledge.

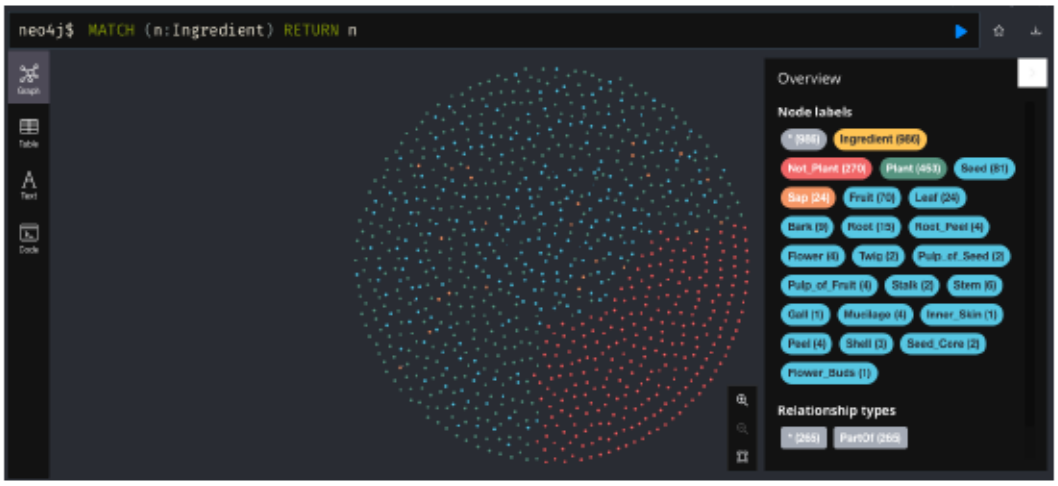


Figure 28 - Snapshot of the user-friendly Neo4j browser for data querying.

[17] Accessible via https://neo4j.com/

Created in 2007, Neo4j became the most widely used graph database framework in the world and is applied in various sectors such as healthcare, government, automotive production. It provides features such as online backups, cache memory, system monitoring, database lock management, and scalability features designed to handle large-scale applications. It supports both embedded and server modes of operation. In embedded mode, Neo4j stores all data on disk which makes it suitable for hardware devices, desktop applications, and embedded server applications.

Furthermore, it uses the Cypher query language that is designed specifically for graph databases which can be directly implemented in its native browser, which helps non-specialists to quickly be able to handle the stored data. It also supports drivers for a wide range of programming languages, including Java, Spring, Scala, and JavaScript and supports exporting query data to JSON and XLS formats, and this facilitates integration into various development environments. This would be relevant to our future research if the Text Mining pipeline was to be rendered in a user-friendly application destined for out-of-domain specialists like historians and biologists.

5.2.2 Data Model

The development of the graph database for old pharmacopeias involved a detailed modeling process. The model (Figure 29) is structured around several fundamental entities such as Remedies, Ingredients, Organs, Symptoms, and Types, Taxa which are provided from the earlier efforts of Named Entity Recognition and Linking. These entities are interconnected through various relationships which represent the structure of the old pharmacopeias that we

analyze. This work was expressed in a preliminary study that was published [121] and further expanded in this chapter.

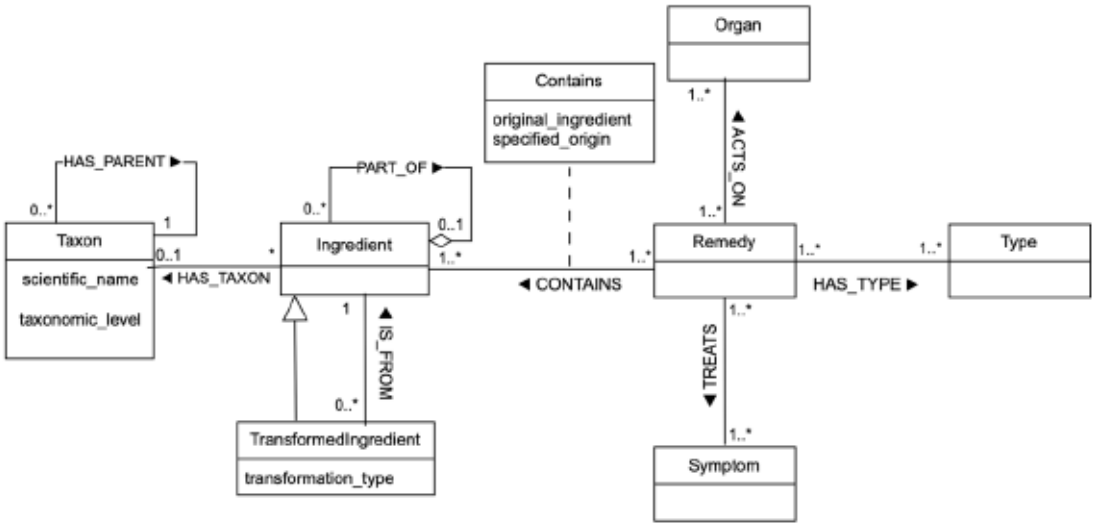


Figure 29 - The database model represented by a UML diagram

The following figures, Figure 30, Figure 31 Figure 32 show a visualization of the results for examples of Cypher queries on Neo4j Browser's interface.

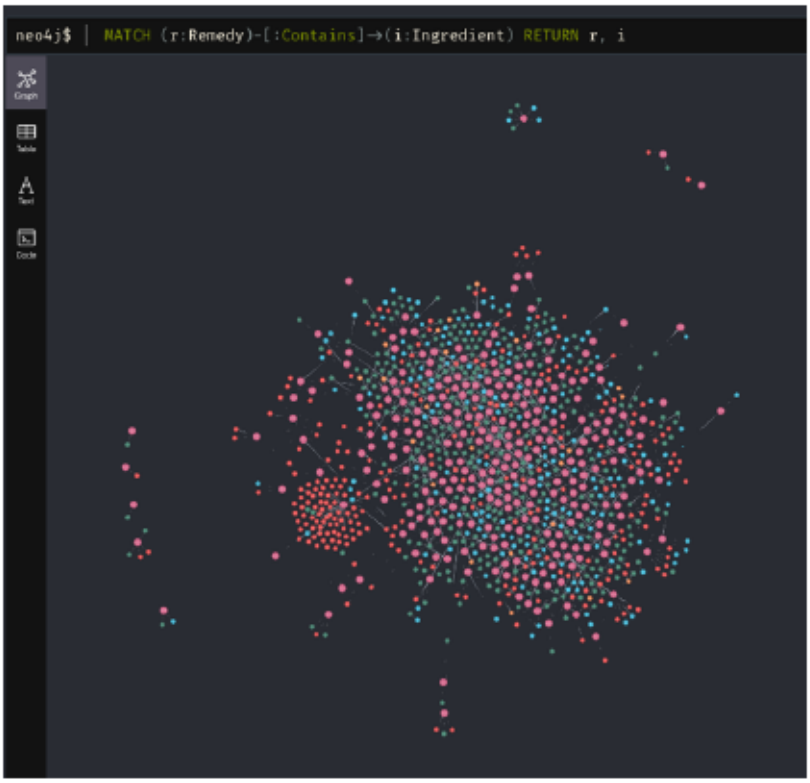


Figure 30 - Cypher query on Neo4j Browser illustrating the different remedies and the ingredients that are contained in them.

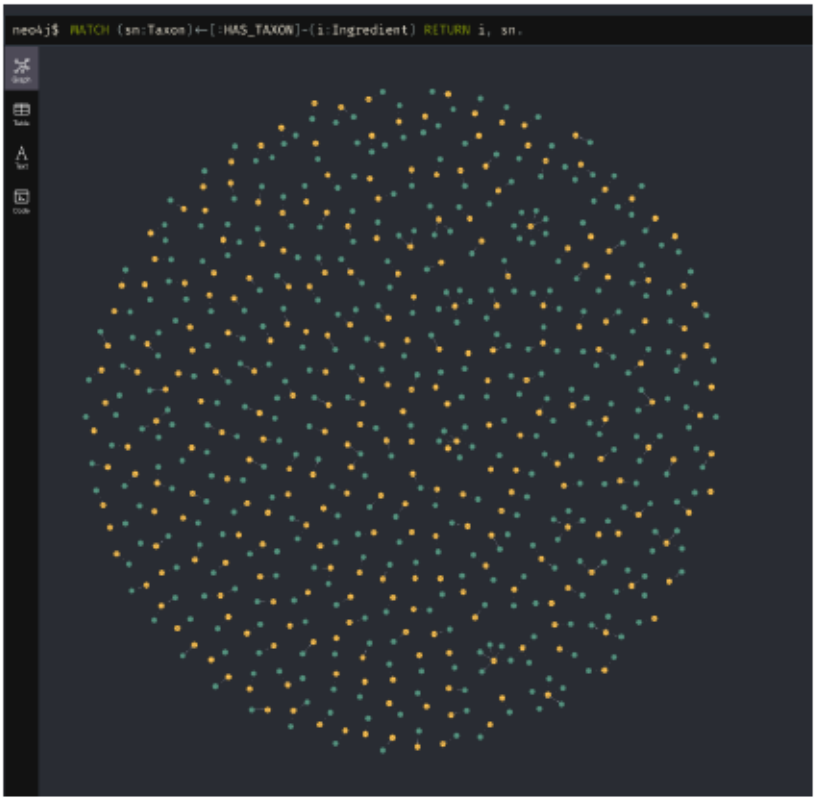


Figure 31 - Cypher query on Neo4j Browser illustrating the different ingredients and the taxa that group them.

Figure 32 - Cypher query on Neo4j Browser illustrating the different remedies that treat the "headache" symptom.

Remedies and Ingredients

The central nodes in the database are the Remedy and Ingredient database model nodes. Each remedy is composed of one or more ingredients, which is taken into account by the CONTAINS relationship. Based on the UML model, in the case of the populated Database for the Sabur Ibn Sahl pharmacopeia, where as a proof of concept we performed the NER annotation manually alongside the manual selection of Taxa for plant-based ingredients, there are 2890 instances of the CONTAINS relationship, linking 292 remedy instance nodes to 986 ingredient instance nodes. Of these ingredients, 716 are plant-based. The CONTAINS relationship includes two attributes:

- The original name of the ingredient as it appears in the manuscript, which was made as a choice in order to preserve the original occurrence before the handling of ingredient structure that is described in the latter Ingredient Parts and Transformations. This was done to keep an accurate backup of the contents of the pharmacopeia for granular and case-by-case studies of the ingredients by the historians or biologists.
- The specified geographic origin of the ingredient when applicable, such as "Antioch" in the case of Antioch scammony. While such ingredients would have similar Taxa as ones that did not have a precise location in their name, this information may be a case of "terroir" that might be of interest to historians when studying remedies.

Ingredient Parts and Transformations

Ingredients in the database can represent an entire plant or a specific part of a plant. To accurately model this, the Ingredient model node is linked by the PART_OF relationship to another Ingredient node. There are 265 database instances of the PART_OF relationship in the Sabur Ibn Sahl pharmacopeia dataset, which reflects the various ways ingredients are derived from different parts of a common plant.

Furthermore, ingredient transformations are an important aspect of the model. Old pharmacopeias describe ingredients that went through processes such as drying, roasting, or boiling. To take this into account, we include the TransformedIngredient model node, which represents the ingredient in its transformed state. Out of the 2890 instances of the CONTAINS relationship, 572 link remedies to transformed ingredients, which are represented by 266 TransformedIngredient database instance nodes. These transformations are further categorized by the IS_FROM relationship (265 instances), which links the transformed ingredient back to its original form and specifies the type of transformation as an attribute.

Organs, Symptoms and Types

The Organ, Symptom and Type classes are also notable components of the UML model. Remedies are associated with specific symptoms (e.g., headaches, fever, abdominal pain) or targeted organs (e.g., stomach, skin), and are of specific types (e.g., beverage, pill, pastille) that capture the therapeutic intentions behind their use in the old pharmacopeias.

For Organ, which has 61 instance nodes in the Sabur Ibn Sahl dataset, the ACTS_ON relationship links remedies to the organs they are designed to treat. This is a many-to-many relationship that shows that a single remedy can target multiple organs. Similarly for Symptom (420 nodes), the TREATS relationship connects remedies to the symptoms they are intended to alleviate. This relationship also operates on a many-to-many basis, as remedies often address multiple symptoms at once. The Type class (represented by 55 database instances) categorizes remedies based on their preparation or administration method, such as pills or decoctions. Each remedy is connected to its type through the HAS_TYPE relationship.

Taxonomic Relationships

Ingredients that are plant-based are connected to a Taxon node via the HAS_TAXON relationship. This relationship indicates the scientific name of the species through the linking of the historical ingredients to modern taxonomic data. In the Sabur Ibn Sahl pharmacopeia, there are 461 instances of the HAS_TAXON relationship, connecting ingredients to their corresponding taxonomic classifications (268 Taxon database instances). This shows how much the data can benefit from structuring through adding taxa, as many different ingredient instances are to be standardized under the Taxon (e.g., both "pomegranate flower" and "pomegranate seeds" are linked to "*punica granatum*").

The taxonomic structure is further detailed by the HAS_PARENT relationship, which links each species to its parent family of species within the broader biological classification. This allows the database to represent the full hierarchy of taxonomic data, with 84 distinct plant families identified in the Sabur Ibn Sahl dataset.

The detailed structure presented earlier in the UML diagram reflects the connections between remedies, ingredients, and their respective taxonomies. For instance, the link between ingredients, where their parts and transformations are represented, extracted from the pharmacopeia and taxonomic data via the HAS_TAXON and HAS_PARENT relationships

demonstrated the model's capacity to join historical and modern scientific data for scientific querying purposes.

5.3 Challenges in Modeling

In the context of the design of the model, the task of representing the ingredients of the old pharmacopeias has presented some challenges. In this section, we will discuss two main aspects: the task of representing plant parts within the graph structure and modeling ingredient transformations

5.3.1 Parts of Plants

Old pharmacopeias tend to often specify not just the plant species but also the particular part of the plant, such as the root, leaf, or bark, that is required to complete the remedies. The distinction is needed between the plant parts and full plants, as different parts of the same plant can have vastly different medicinal properties.

To address this need, we made the database model incorporate a PART_OF relationship that links an Ingredient class to another Ingredient representing a specific plant part. This relationship allows for the hierarchical structuring of plant-based ingredients, where a broader Ingredient (e.g., vernacular plant name) can be connected to multiple model nodes representing its parts (e.g., roots, leaves, flowers).

If the original ingredient is "apple seeds" the actual used ingredient that is linked to the Remedy is to be indicated as such: "seed_fruit_apple tree". The ingredient "apple seeds" the chain goes from "Seed: seed_fruit_apple tree" to the plant "Plant: apple tree" passing through "Fruit: fruit_apple tree"." This illustrates three categories of plant parts: the seed, the fruit, and the tree (the entire plant). Each part of a plant has been modeled as a node in a hierarchical tree (Figure 33). The root of the tree is the "entire plant" node, which is linked to the corresponding taxon in our model.

Plant sub-parts have been grouped into 21 types from examining Sabur Ibn Sahl's pharmacopeia: Fruit, Pulp of fruit (inner layer of a fruit, often thin and fibrous), Inner skin (inner skin of a fruit), Peel (outer layer of a fruit), Shell (hard outer layer of certain fruits), Seed, Pulp of seed, Seed core, Seed vessel (seed cover), Stem, Leaf, Twig, Stalk (stem that supports the flower and then the fruit), Flower, Flower buds, Root, Root peel (outer layer of the root), Bark, Mucilage (thick, viscous substance produced by certain plants), Sap, Gall (abnormal plant growth caused by bacterial infection).

It is notable to add that for each Ingredient instance that is plant-based, in order to facilitate querying, an additional label was given for each node as to specify the part of plant from these 21 types, a "whole plant" label, as well as a "not_plant" label for ingredients that are from other sources like animal or mineral.

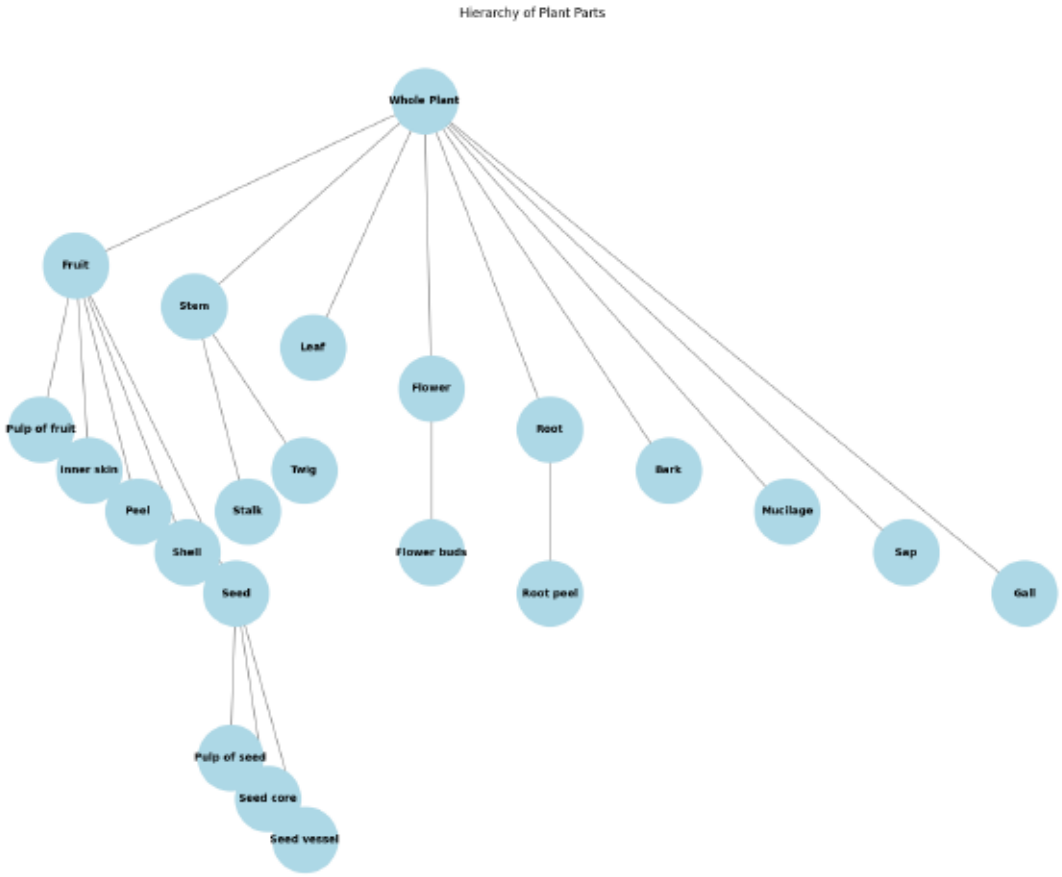


Figure 33 - Hierarchy of the plant parts.

Each ingredient is broken down according to the chain of sub-parts that constitute it. A Python script was used to populate the database according to the model, using the hierarchy of plant sub-parts[18]. This allowed for a more precise modeling of plant parts and linking remedies to the plants from which their ingredients are derived which in turn would help with the retrieval of information. For example, Figure 34 shows the representation of the relationship of the citron fruit which is contained in remedies 41 and 252 of a given pharmacopeia, while citron

[18] Python code for Database population accessible via https://github.com/karimhaff/pharmacopeias/blob/main/db_populate

peels are contained in remedy 252. If a remedy requires the peel of the citron, the database will include an Ingredient node for the citron fruit and a linked Ingredient node specifically for the peel of that citron fruit. The PART_OF relationship between these nodes indicates that the peel is a sub-part of the broader citron fruit which is a part of the broader citron tree. This approach ensures that the database reflects the pharmacopeia's instructions of which ingredients are really used because we have grouped and standardized plant ingredients; in the case of the citron, the grouping of all the parts related to it aids in making them appear in the query results and permit querying on many levels of the plant-part hierarchy.

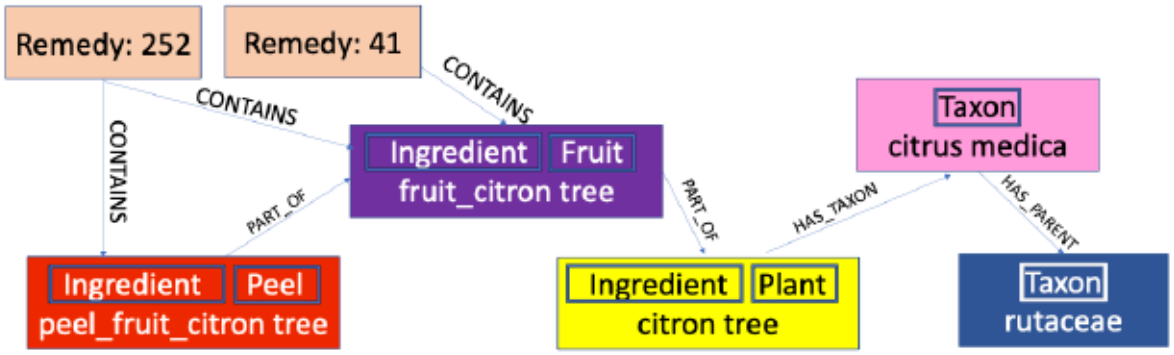


Figure 34 - Representation of remedies which contain ingredients that are parts of the citron tree.

This modeling choice is important when considering the diverse medicinal applications of different plant parts. In many cases, the parts of any plant may have distinct uses and medicinal effects. Explicitly representing these parts allow the subsequent tasks of insight extraction to be able to study correlations between specific parts and their medicinal applications.

However, modeling plant parts also presents difficulties, specifically in ensuring that the relationships between different parts and the whole plant are accurately represented because the hierarchical nature of these relationships can become complex if scientific plant anatomy ontologies that contain highly specialized vocabulary were to be used for this task. In our case, we created the hierarchy based on manual census upon examining Sabur Ibn Sahl's pharmacopeia. This manual approach, while effective for our specific needs, sheds light unto the difficulty of mapping these parts to extensive plant anatomy ontologies. Many plant parts mentioned in vernacular language are not easily placed within modern anatomical frameworks due to a lack of expertise of the computational linguist in botanical matters while designing database model for Text Mining pipeline. Addressing this difficulty would require

collaboration with a botanist or a specialist in plant taxonomy who could provide the necessary knowledge to map these terms to a more accurate anatomical framework.

Further, the task of creating an automated rule-based algorithm to recognize and categorize the plant parts is complicated by the limitations of current NLP techniques when applied to pharmacopeias. Acknowledging that it is possible to use dictionaries and rule-based approaches to detect explicitly mentioned plant parts in the text (e.g., liquorice **roots**), the task becomes more difficult to automate when dealing with implicit references. For example, a term like "olive" refers to the fruit of the olive tree, but the word "fruit" is not explicitly mentioned in order to automatically place as a fruit within the hierarchy. Additionally, the same can be said about the lack of the explicit mention of the olive tree that dictates the relationship between the part (olive) and the whole plant (olive tree) by that it is linked to the Taxon node. This type of implicit information is difficult to extract automatically and would require a deeper level of semantic analysis that goes beyond literal word matching.

5.3.2 Ingredient Transformations

Old pharmacopeias often detail procedures that convert raw ingredients into different forms such as drying, roasting, or grinding each of which may alter the ingredient's medicinal properties. Our goal was to represent both the original ingredient and its transformed states.

In the model, transformations are handled through the TransformedIngredient class. This model node is linked to the original Ingredient via the IS_FROM relationship, which explicitly indicates that the transformed substance is derived from the original ingredient. The transformation type is recorded as an attribute within the TransformedIngredient node, which provides a categorical classification of the transformation without providing a specific hierarchy between the different transformations as we did in Figure 33, as we consider that there is no obvious hierarchy such as with plant parts. This allows for a clear distinction between the raw and altered forms of an ingredient, preserving the integrity of the historical data. Table 9 showcases the different possible transformation types that were found in Sabur Ibn Sahl's pharmacopeia that we group here for clarity.

Category	Transformation Type	Example
Drying, Burning and Preservation	Drying	dried violet
	Fermentation	aged wine

	Roasting	roasted coriander
	Ash Creation	ashes of vine stems
	Burning	burnt staghorn
Physical Alterations	Skin/Seed Removal	peeled seeds of the musk melon, seedless raisins
	Core/Stem/Berry Removal	coreless olives, stemless rose, berryless Syrian sumach
	Stalk Removal	stalkless red roses
	Fiber Removal	Fibreless tamarinds
	Grinding/Chopping	ground tragacanth, chopped quinces
	Scraping	scraped liquorice roots
Liquids, Boiling and Cooking	Infusion	infused wine
	Juicing	juice of the soft-rinded pomegranate
	Beverages	myrtle beverage
	Water-Based	rose-water
	Oil-Based	oil from unripe olives
	Boiling	boiled wine
	Cooking	cooked anise-water
Barking and Husk Removal	Husk Removal	husked bitter almonds
	Bark Removal	barked fig-wood
Natural Transformation	Ripeness/non-ripeness	ripe succulent myrtle berries, unripe olives

Table 9 - The different transformation types.

For instance, consider a remedy that requires seedless barberries. In the database, the Ingredient node for barbery would be linked to a TransformedIngredient node labeled as seedless_fruit_barberry as the transformed plant part is the fruit. The IS_FROM relationship between these two nodes makes it clear that the seedless barberry is a transformation of the original raw ingredient (Figure 35). This approach ensures that the transformation process is explicitly documented within the database, allowing researchers to trace the lineage of each ingredient and understand how its preparation might affect its medicinal use.

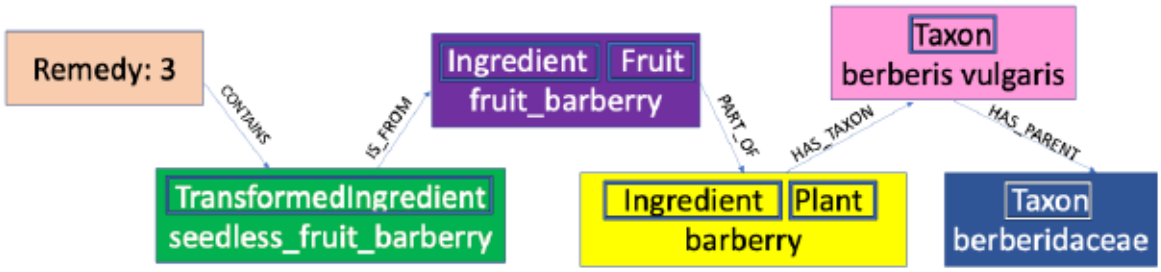


Figure 35 - Representation of a remedy which contains a transformed part of the barberry plant.

Despite its advantages, this approach also has some difficulties. One issue is the potential proliferation of TransformedIngredient nodes such as for ingredients that can undergo a wide range of transformations. Managing these nodes and ensuring that they are accurately linked to their corresponding original ingredients requires a knowledge of chronology between different transformations applied to the same ingredient. Further, the reliance on manual entry for these transformation types due to the limitations of an automated process that cannot exhaustively take into account the rich vocabulary of transformation in natural language introduces additional complexity, as it necessitates perpetual updates as more pharmacopeias are studied and new transformation types are found.

5.4 Exploring Remedies with Formal and Relational Concept Analysis

5.4.1 Ingredient Frequency and Co-occurrence with Formal Concept Analysis

As presented and introduced in 2.3.4, Formal Concept Analysis, or FCA, is a mathematical approach used to derive a conceptual structure from a dataset, represented as a formal context. In the context of our Text Mining pipeline, FCA was employed to analyze the relationships between remedies and their constituent plant-based ingredients in order to identify the most frequent and co-occurring ingredients in Sabur Ibn Sahl's pharmacopeias. Our work relating to this is the main topic of a publication [122].

The data model utilized for this section consists of a formal context that describes 287 remedies (rows) and their 586 plant-based ingredients (columns) that were extracted from the graph database through a Python code that was developed to convert the JSON data to an RCFT file format[19]. This conversion was done in order to fit the input requirements of the RCAExplore tool which is specialized in the task at hand [123]. We refer to this dataset as *Dataset_1* and its goal is to obtain answers to questions concerning the frequency and co-occurrence of ingredients. Table 10 presents an excerpt from this formal context and illustrates how the data is organized.

remedies/ingredients	alhagi	asarabacca	barberry	barley	camphor tree	fruit_barberry	...
Remedy: 1	×			×			
Remedy: 2					×	×	
Remedy: 3	×		×		×	×	
Remedy: 4	×	×	×			×	
...							

Table 10 - A sample extract from the remedies-ingredients context.

The analysis of *Dataset_1* using FCA resulted in a concept lattice containing 1,158 concepts, excluding the top and bottom elements of the lattice. This lattice provides a hierarchical

[19] Python code accessible via https://github.com/karimhaff/pharmacopeias/blob/main/rcft_converter

representation which aids in seeing what ingredients are shared by the remedies. The concepts within this lattice range from the most general, which include the most frequent ingredients, to the most specific, which may include only one or a few remedies. A top-down traversal of the lattice was conducted to investigate the most frequent ingredients and their associated remedy clusters. This exploration is relevant for addressing the research question related to the most commonly used ingredients in the pharmacopeia. For instance, one of the most general concepts (summarized in Table 11) identified in this analysis is represented by the ingredient "plant for wine" which appears in 51 remedies. 51 of the 287 remedies have wine or wine vinegar as ingredient, probably used as thinner, and generally without precision on the original plant, grapes or others[20].

One ingredient in concept intent	- {plant for wine} (51) - {saffron} (33) -{seed_sesamum} (29) - {bark_cinnamon tree} (23)	- {rose} (47) - {sap_acacia (34) - {indian spikenard} (32) - {mastic} (32) - {plant of vinegar} (23) - {ginger} (23) - {sap_tragacanth} (22 -{fruit_olive tree} (21) ...
Two ingredients in concept intent	-{plant for wine, saffron} (16) -{sap_tragacanth, sap_acacia} (14) -{ginger, bark_cinnamon tree} (13) -{plant for wine, rose} (11) ...	- {saffron, indian spikenard} (14) -{plant for wine, indian spikenard} (13) - {mastic, indian spikenard} (12)
Three ingredients in concept intent	-{ginger, long pepper, black pepper} (9) -{ginger, bark_cinnamon tree, long pepper} (8) -{saffron, ginger, bark_cinnamon tree} (7) -{clove, ginger, long pepper} (7) ...	- {saffron, plant for wine, bark_cinnamon tree } (9) - {indian spikenard, long pepper, black pepper} (8) -{plant for wine, myrrh, bark_cinnamon tree} (6)

Table 11 - Summary of the most general concepts their extent cardinality in brackets, and the detail of their intent (ingredients).

This ingredient is followed by "rose," which is present in 47 remedies. These frequent ingredients likely played important roles in the formulation of remedies, either as primary components or as common additives.

In addition to identifying frequent ingredients, the FCA analysis also revealed co-occurring ingredients, sets of ingredients that frequently appear together in remedies. These co-occurring ingredients are of interest as they may indicate synergistic relationships in the preparation of remedies. As the concept lattice is very large, a reliable way to search for the co-occurring ingredients is to generate the base of implications. Table 12 summarizes the results of the

[20] Every wine-related ingredient where the original plant was not specified was standardized in the form of the plant name "plant for wine".

implication rules obtained from *Dataset_1* and indicates the number of rules per support within 287 remedies.

Table 12 - Implication rule summary

Number of rules	Support
895	2
273	3
115	4
25	5
16	6
5	7
1	9

Rule	Support
liquorice, sap_acacia → sap_tragacanth	9
black pepper, ginger, indian spikenard → long pepper	7
cassia, myrrh, plant for wine → saffron	7
bark_cinnamon tree, cassia → saffron	6
mace → clove	6
seed_quince → sap_acacia	5

Table 13 – Implication Examples from Dataset_1.

Table 13 shows common implications rules. For instance, if an implication rule in the table reads liquorice, sap_acacia → sap_tragacanth with a support of 9, it means that in the dataset, whenever both liquorice and sap_acacia are present in a remedy, sap_tragacanth is also present in 9 remedies. Indeed, this helps to identify strong associations between ingredients in the remedies. The higher the support, the more frequently the combination occurs in the dataset. These rules can be used to infer potential ingredient combinations that were commonly used together in old pharmacopeias which is beneficial for the investigation of insights.

5.4.2 Application of Relational Concept Analysis

Relational Concept Analysis, or RCA, is an extension of FCA that allows for the incorporation of relational data into the conceptual analysis. RCA is relevant when the relationships between different types of objects need to be considered in addition to the attributes of those objects. In our work, RCA was employed to study the relationships between remedies and ingredients further as to answer the question "what families of ingredients are mostly associated to a specific group of symptoms?".

To address this question, a Relational Concept Family (RCF) model, termed *Dataset_2*, was constructed to analyze the relationships between symptoms, remedies, and ingredients. This model is illustrated in Figure 36.

Figure 36 - Model for the Symptoms-Remedies-Ingredients RCF.

The RCF model consists of three formal contexts: Symptoms (with a Category as an attribute), Remedies (with a From as an attribute), and Ingredients (with a Taxon as an attribute), along with two relational contexts: *isTreatedBy*, which links symptoms to remedies, and *isComposedOf*, which links remedies to their ingredients. For the analysis, data is restricted to a subset of remedies treating fever symptoms, their ingredients and symptoms. Names of symptoms are diverse and were manually categorized into 8 categories that we focus on during the analysis. These categories are ranked from the most frequent to the less frequent, specifying in brackets the number of symptoms: fever (19), digestive (12), dermatological (12), neurological (12), hepatic (6), psychiatric (5), respiratory (4), hematological (3). Indeed, the *Symptoms* context categorizes symptoms by type, the *Remedies* context describes 26 remedies by their forms, and the *Ingredients* context classifies 156 ingredients by their taxonomic details, such as species and family.

As shown in Figure 37, the *remedies* lattice contains 13 concepts. Pastille is the most common form of remedy (C_remedies_13) followed by potion (C_remedies_12). Interestingly, some remedies appear in different forms, such as remedies 1 and 5 in C_remedies_9 which have pastille and potion forms. In both of them, a pastille is prepared and then dissolved before consumption to make a potion. This also shows potential insight about not only the contents of a remedy but also its mode of preparation which would be of interest for future investigation.

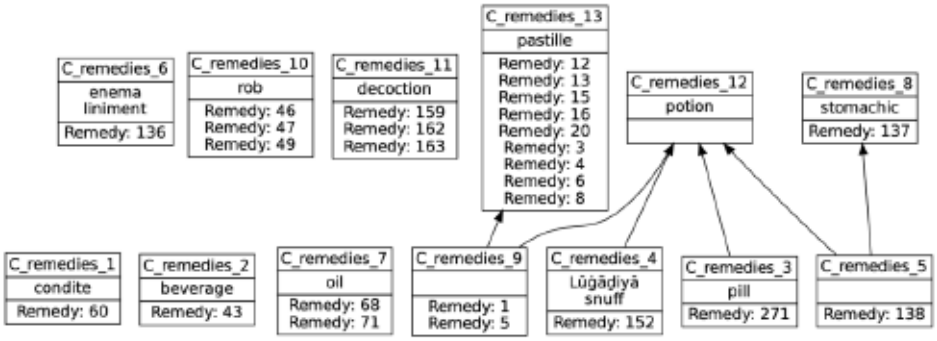


Figure 37 - The Remedies lattice from Dataset_2 (without top and bottom elements).

Further, the Ingredients lattice contains 121 concepts and groups ingredients by their family taxon and plant species taxon. Grouping concepts of ingredients by family is of interest because plants of the same family may share properties. Figure 38 shows a sample from the Ingredients lattice which sheds light onto the two of most frequent ingredient families: apiaceae, 14 ingredients (C_ingredients_121) and asteraceae, with 10 ingredients (C_ingredients_120), that

seem to be the families of plants that are most closely related to the treatment of an illness that causes fever.

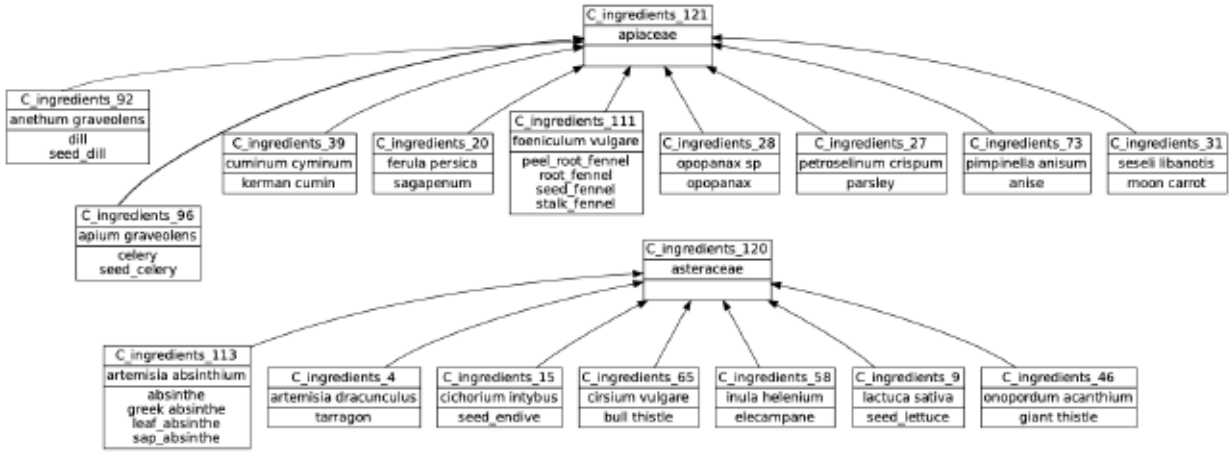


Figure 38 - Sample of the Ingredients lattice from Dataset_2.

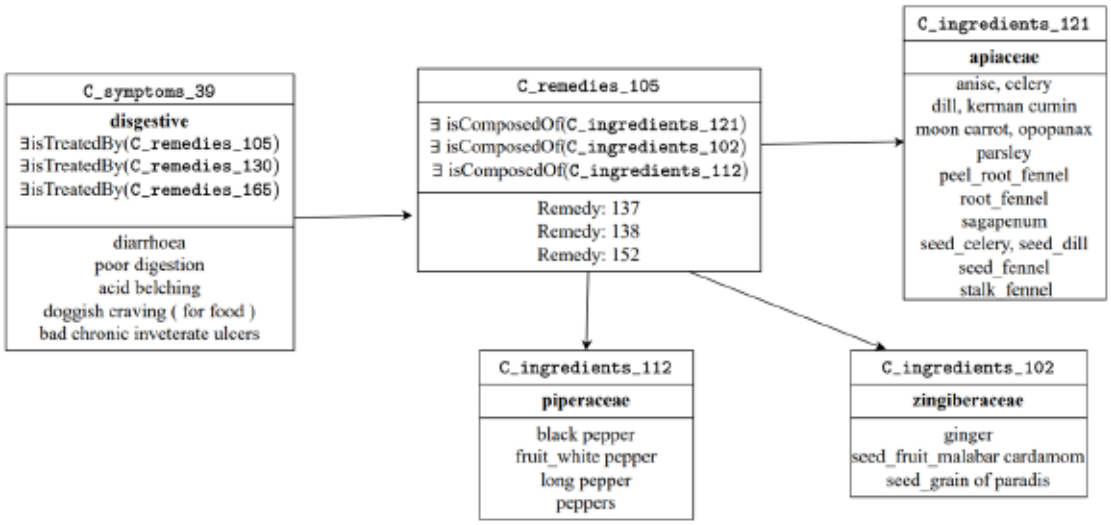


Figure 39 - An extract from RCA results on Dataset_2 (with Existential quantifiers-[124]).

Further, Figure 39 illustrates a sample from a RCA which relates ingredients (classed by family), the remedies they are contained in and associated symptoms. It provides an insight that the digestive symptoms mentioned may be treated using plants belonging to *apiaceae*, *zingiberaceae* and *piperaceae* families. This example demonstrates how RCA can be used to answer specific queries about the data, such as "what families of ingredients are mostly

associated to a specific group of symptoms?". Indeed, all in all, the use of FCA and RCA helps to perform a wide analysis of the stored data that we have from old pharmacopeias. This analysis can bring about insights from spotted patterns and relationships and these insights can inform future research and support the continued exploration of old medical texts, as a finality of the Text Mining pipeline.

5.4.3 Conclusion

The development of an information model for old pharmacopeias, as detailed in this chapter, has shed light onto the complexity in representing the various relationships between the many entities present in such documents.

The decision to employ a graph database proved beneficial in handling these complexities through its ability to model interconnected data and accommodate the hierarchical nature of plant parts and ingredient transformations. Though, several challenges emerged such as in the modeling of complex ingredient structures and the limitations posed by manual data entry and manual tagging.

Further, the application of FCA and RCA offered a beneficial method for querying data stored in the graph database and aided in revealing patterns within the data. These methods provided a structured means of answering research questions, such as those related to the co-occurrence of ingredients and their association with specific symptoms, which will be the basis upon which future research will be conducted in search of new ways to design medical drugs. Thus, going forward, collaboration with domain experts will be necessary to fully realize the potential of this approach in the study of old pharmacopeias.

Chapter 6: General Conclusion

Throughout this thesis, we addressed the primary difficulty of transforming unstructured, pharmacopeic textual data into structured and analyzable formats. This was done, informed by the literature, in the form of a Text Mining pipeline to valorize the contents of pharmacopeias and make use of extracted insights that can potentially lead to advancements in the search for antibiotic alternatives.

For the initial phase of the research, in Chapter 3: we focused on building and assessing a NER resource that is specific to the content of pharmacopoeias through using specialized tags (organs, symptoms, ingredients and remedy types). This involved fine-tuning and comparing several transformer-based language models to recognize and categorize the named entities. The NER system that used the best performing model (DeBERTaV3 [11]) demonstrated solid performances. The main substance of this contribution is the ability to automate a process that would traditionally require extensive manual effort, as well as publishing a resource that can be used by the community.

Following the implementation of the NER system, we advanced to a second contribution in Chapter 4: regarding the challenge of Named Entity Linking and Disambiguation. The goal here was to accurately link the named entities recognized in the texts, namely plant names, to modern taxonomic identifiers in the data infrastructure GBIF. At first, this task was deemed challenging by the variability in plant nomenclature which is ambiguous when written vernacularly. To overcome these challenges, we created a system of candidate selection based on geo-reference to make way for an adequate identification of plant species based on their geographical occurrence relative to the location where the pharmacopoeia was authored. We concluded that this task was met with data losses caused by the lack of entries in the data infrastructures that were used in the pipeline.

In the third phase which is Chapter 5: we described the structuring of the extracted data from the first and second phases within a graph database. This approach aimed to adequately represent complex relationships between the various named entities that lie with the pharmacopeias that we have worked on. The goal of this Information Modeling was to create a base from which insights could be extracted from these pharmacopeias. Formal Concept Analysis and Relational Concept Analysis were the methods that we used to extract insights. This analysis showed the potential the lies within querying old pharmacopeias and

demonstrated the ability to uncover idea such as the prevalence of plants from the plant family of *apiaceae* which was prevalent in remedies for digestive symptoms within the manuscript of Sabur Ibn Sahl.

Additionally, a work in progress by the pharmacognosy team demonstrated potential anti-Mycobacterial activity (Figure 40) from substances that we found in Sabur Ibn Sahl's pharmacopeia in a liniment that is described to heal the tuberculin disease of Scrofula (Remedy 264). Initial experimental testing of these ingredients in isolation revealed that the common base ingredients, except for copper, did not exhibit activity against the target bacteria, Mycobacterium tuberculosis. However, the newly identified plants based on a preliminary Neo4j query of co-occurrence relating the ingredients of Remedy 264 to their most common co-occurents in other remedies (in this case, opoponax, bdellium and galbanum) demonstrated a modest yet promising antibacterial activity. As this is a work in progress, future findings are to come, and we hypothesize that the future use of the Text Mining pipeline that we have developed in this thesis may aid in finding more of such experiments to engage in, going forward.

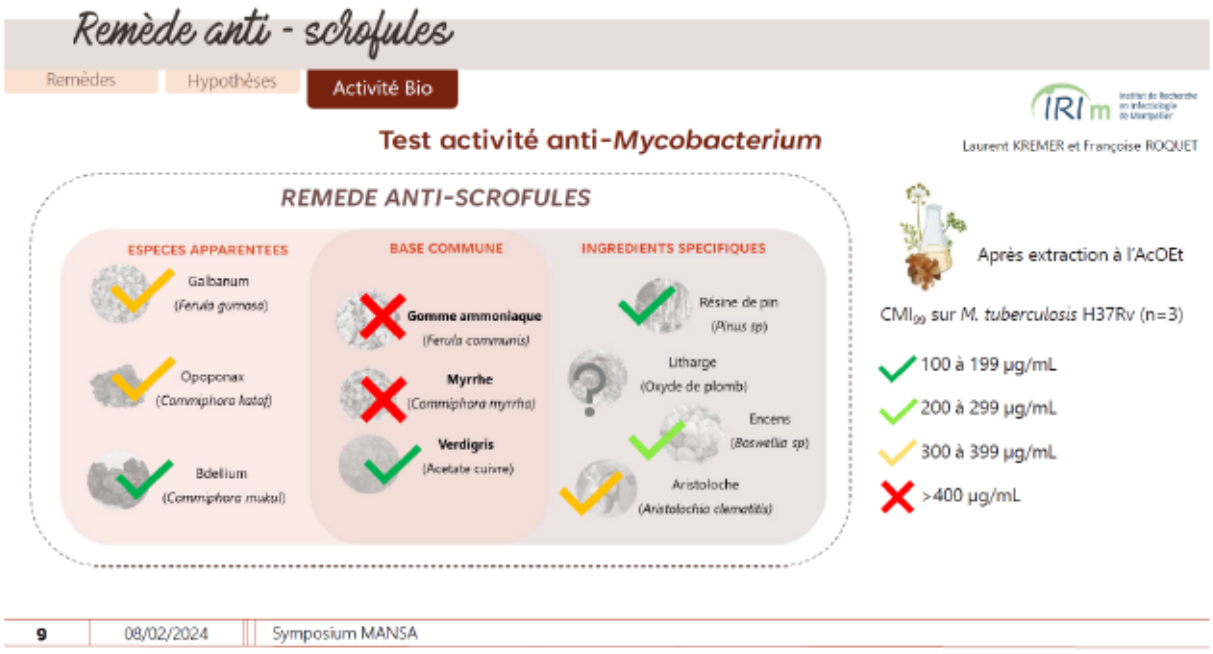


Figure 40 – Findings of an underway project regarding the Scrofula disease © Capucine Braillon.

Indeed, we emphasize the importance of interdisciplinary collaboration in addressing complex research questions. The integration of historical knowledge, textual analysis, and medical

expertise in one collaboratory framework has demonstrated an important utility for our research.

Looking forward, the methodologies and findings of this thesis reveal many aspects for further investigation in regard to Text Mining. Alongside future improvements to the many tasks of the Text Mining pipeline as we have discussed throughout the thesis, one new potential direction is the application of the developed techniques to other pharmacopeias, such as those from different authors or even different heritages that were translated to English as it is the main language being processed in this work. Indeed, given the translated nature of the pharmacopeias that we work on, future work could also explore the use of transformer models for the Arabic language such as AraBERT [125] to test performance on the original Arabic texts of the same manuscripts and comparing the results with its English counterpart in order to study the generalizability of the NER task on a less-resourced language, and experiment by reusing multilingual resources such as the BabelNet to test its ability to disambiguate Arabic named entities.

As for the geographical-based candidate selection in NEL, to account for a possible limited historicity of considering only a radius around Baghdad or around the origin location of each pharmacopeia, it would be pertinent to take into account candidate plants present in other radii that can modify the scores (such as Greece due to some plagiarism of Abbasid authors, China or India due to trade routes). This would require modifying the calculation parameters. This can be done within the developed method, by defining new radii and coordinates on the map (with importance weights) and rerunning the same calculation to obtain the new scores. Defining these "zones of influence" can be done in direct collaboration with historians.

We also consider further extension of the developed database in collaboration for specialists such as botanists, in order to account for the limited nature of the resources that we have used in this work.

The work also may open the door of analyzing Kitāb al-bayān fī kashf ʿasrār al-ṭibb lil-ʿiyān (The Unveiling of the Secrets of Medicine for All) by Al Ḥamawī, from the 13th century [6], which is the pharmacopeia that first raised the idea of this research.

References

[1] Y. Houdas, 'La médecine arabe aux siècles d'or VIIè-XIIIè siècle', pp. 1–164, 2003.

[2] F. Sanagustin, 'Manfred Ullman, La médecine islamique. Trad, de l'anglais par F. Hareau, PUF, Coll. Islamiques, Paris, 1995', 1999, Accessed: Aug. 28, 2024. [Online]. Available: https://www.persee.fr/doc/bcai_0259-7373_1999_num_15_1_977_t1_0175_0000_2

[3] F. Harrison, A. E. L. Roberts, R. Gabrilska, K. P. Rumbaugh, C. Lee, and S. P. Diggle, 'A 1,000-Year-Old Antimicrobial Remedy with Antistaphylococcal Activity', *mBio*, vol. 6, no. 4, pp. e01129-15, Aug. 2015, doi: 10.1128/mBio.01129-15.

[4] E. Connelly, C. I. del Genio, and F. Harrison, 'Data Mining a Medieval Medical Text Reveals Patterns in Ingredient Choice That Reflect Biological Activity against Infectious Agents', *mBio*, vol. 11, no. 1, pp. e03136-19, Feb. 2020, doi: 10.1128/mBio.03136-19.

[5] X. Su and L. H. Miller, 'The discovery of artemisinin and Nobel Prize in Physiology or Medicine', *Sci. China Life Sci.*, vol. 58, no. 11, pp. 1175–1179, Nov. 2015, doi: 10.1007/s11427-015-4948-7.

[6] E. Oussiali and V. Pitchon, 'Un traité singulier de médecine arabe médiévale : identification d'un manuscrit', *Rev. BNU*, no. 22, Art. no. 22, Nov. 2020, doi: 10.4000/rbnu.4802.

[7] T. Al-Moslmi, M. Gallofré Ocaña, A. Opdahl, and C. Veres, 'Named Entity Extraction for Knowledge Graphs: A Literature Overview', *IEEE Access*, vol. 8, pp. 32862–32881, Feb. 2020, doi: 10.1109/ACCESS.2020.2973928.

[8] V. S. Pendyala, Y. Fang, J. Holliday, and A. Zalzala, 'A text mining approach to automated healthcare for the masses', in *IEEE Global Humanitarian Technology Conference (GHTC 2014)*, Oct. 2014, pp. 28–35. doi: 10.1109/GHTC.2014.6970257.

[9] R.-G. Yu *et al.*, 'Text Mining-Based Drug Discovery in Osteoarthritis', *J. Healthc. Eng.*, vol. 2021, pp. 1–14, Apr. 2021, doi: 10.1155/2021/6674744.

[10] S. Brumfield, 'Imperial Methods: Using Text Mining and Social Network Analysis to Detect Regional Strategies in the Akkadian Empire', Ph.D., Ann Arbor, United States, 2013. Accessed: Jun. 12, 2023. [Online]. Available: https://www.proquest.com/docview/1400836723/abstract/C50CD928ACFB490BPQ/1

[11] P. He, X. Liu, J. Gao, and W. Chen, 'DeBERTa: Decoding-enhanced BERT with Disentangled Attention', Oct. 06, 2021, *arXiv*: arXiv:2006.03654. doi: 10.48550/arXiv.2006.03654.

[12] A. Hamdi *et al.*, 'A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers', in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, in SIGIR '21. New York, NY, USA: Association for Computing Machinery, Jul. 2021, pp. 2328–2334. doi: 10.1145/3404835.3463255.

[13] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, and A. Doucet, 'Named Entity Recognition and Classification in Historical Documents: A Survey', *ACM Comput. Surv.*, vol. 56, no. 2, pp. 1–47, Feb. 2024, doi: 10.1145/3604931.

[14] M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, and S. Clematide, 'Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents', in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G.

Pasi, A. Hanbury, M. Potthast, G. Faggioli, and N. Ferro, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, pp. 423–446. doi: 10.1007/978-3-031-13643-6_26.

[15] C. M. Phan, 'Named entity recognition and linking with knowledge base', Nanyang Technological University, 2019. doi: 10.32657/10356/136585.

[16] J. Lee *et al.*, 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining', *Bioinformatics*, p. btz682, Sep. 2019, doi: 10.1093/bioinformatics/btz682.

[17] E. Boros *et al.*, 'Alleviating Digitization Errors in Named Entity Recognition for Historical Documents', in *Proceedings of the 24th Conference on Computational Natural Language Learning*, R. Fernández and T. Linzen, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 431–441. doi: 10.18653/v1/2020.conll-1.35.

[18] S. Malmasi, A. Fang, B. Fetahu, S. Kar, and O. Rokhlenko, 'SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER)', in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, Eds., Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1412–1437. doi: 10.18653/v1/2022.semeval-1.196.

[19] J. Su *et al.*, 'Global Pointer: Novel Efficient Span-based Approach for Named Entity Recognition', Aug. 05, 2022, *arXiv*: arXiv:2208.03054. doi: 10.48550/arXiv.2208.03054.

[20] M. Ehrmann, M. Romanello, A. Flückiger, and S. Clematide, 'Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers', in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, and N. Ferro, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 288–310. doi: 10.1007/978-3-030-58219-7_21.

[21] K. El Haff, W. Antoun, F. Le Ber, and V. Pitchon, 'Reconnaissance des entités nommées pour l'analyse des pharmacopées médiévales', in *EGC 2023 - Extraction et Gestion des Connaissances*, Lyon, France, 2023. [Online]. Available: https://hal.science/hal-03934557

[22] A. Sitter, T. Calders, and W. Daelemans, 'A Formal Framework for Evaluation of Information Extraction', Jul. 2004.

[23] P. J. Gorinski *et al.*, 'Named Entity Recognition for Electronic Health Records: A Comparison of Rule-based and Machine Learning Approaches', Jun. 05, 2019, *arXiv*: arXiv:1903.03985. doi: 10.48550/arXiv.1903.03985.

[24] B. Alex, C. Grover, R. Tobin, C. Sudlow, G. Mair, and W. Whiteley, 'Text mining brain imaging reports', *J. Biomed. Semant.*, vol. 10, no. 1, p. 23, Nov. 2019, doi: 10.1186/s13326-019-0211-7.

[25] E. Kogkitsidou and P. Gambette, 'Normalisation of 16th and 17th century texts in French and geographical named entity recognition', in *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, in GeoHumanities'20. New York, NY, USA: Association for Computing Machinery, Nov. 2020, pp. 28–34. doi: 10.1145/3423337.3429437.

[26] D. Küçük and A. Yazıcı, 'Named Entity Recognition Experiments on Turkish Texts', in *Flexible Query Answering Systems*, T. Andreasen, R. R. Yager, H. Bulskov, H.

Christiansen, and H. L. Larsen, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2009, pp. 524–535. doi: 10.1007/978-3-642-04957-6_45.

[27] P. Thompson, J. McNaught, and S. Ananiadou, 'Customised OCR correction for historical medical text', in *2015 Digital Heritage*, Sep. 2015, pp. 35–42. doi: 10.1109/DigitalHeritage.2015.7413829.

[28] K. McDonough, L. Moncla, and M. van de Camp, 'Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora', *Int. J. Geogr. Inf. Sci.*, vol. 33, no. 12, pp. 2498–2522, Dec. 2019, doi: 10.1080/13658816.2019.1620235.

[29] K. Kettunen, E. Mäkelä, T. Ruokolainen, J. Kuokkala, and L. Löfberg, 'Old Content and Modern Tools - Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771-1910', Nov. 09, 2016, *arXiv*: arXiv:1611.02839. doi: 10.48550/arXiv.1611.02839.

[30] C. Wen, T. Chen, X. Jia, and J. Zhu, 'Medical Named Entity Recognition from Un-labelled Medical Records based on Pre-trained Language Models and Domain Dictionary', *Data Intell.*, vol. 3, no. 3, pp. 402–417, Sep. 2021, doi: 10.1162/dint_a_00105.

[31] P. D. Soomro, S. Kumar, Banbhrani, A. A. Shaikh, and H. Raj, 'Bio-NER: Biomedical Named Entity Recognition using Rule-Based and Statistical Learners', *Int. J. Adv. Comput. Sci. Appl. Ijacsa*, vol. 8, no. 12, Art. no. 12, Jul. 2017, doi: 10.14569/IJACSA.2017.081220.

[32] S. Morwal, 'Named Entity Recognition using Hidden Markov Model (HMM)', *Int. J. Nat. Lang. Comput.*, vol. 1, no. 4, pp. 15–23, Dec. 2012, doi: 10.5121/ijnlc.2012.1402.

[33] A. Ekbal and S. Bandyopadhyay, 'A Hidden Markov Model Based Named Entity Recognition System: Bengali and Hindi as Case Studies', in *Pattern Recognition and Machine Intelligence*, vol. 4815, A. Ghosh, R. K. De, and S. K. Pal, Eds., in Lecture Notes in Computer Science, vol. 4815. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 545–552. doi: 10.1007/978-3-540-77046-6_67.

[34] S. Baluja, V. O. Mittal, and R. Sukthankar, 'Applying Machine Learning for High-Performance Named-Entity Extraction', *Comput. Intell.*, vol. 16, no. 4, pp. 586–595, Nov. 2000, doi: 10.1111/0824-7935.00129.

[35] H. Isozaki, 'Japanese Named Entity Recognition based on a Simple Rule Generator and Decision Tree Learning', in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France: Association for Computational Linguistics, Jul. 2001, pp. 314–321. doi: 10.3115/1073012.1073053.

[36] O. Bender, F. J. Och, and H. Ney, 'Maximum Entropy Models for Named Entity Recognition', in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 148–151. Accessed: May 20, 2024. [Online]. Available: https://aclanthology.org/W03-0420

[37] J. Kazama, T. Makino, Y. Ohta, and J. Tsujii, 'Tuning support vector machines for biomedical named entity recognition', in *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*, Phildadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 1–8. doi: 10.3115/1118149.1118150.

[38] A. Ekbal, R. Haque, and S. Bandyopadhyay, 'Named Entity Recognition in Bengali: A Conditional Random Field Approach', in *Proceedings of the Third International Joint*

Conference on Natural Language Processing: Volume-II, 2008. Accessed: May 20, 2024. [Online]. Available: https://aclanthology.org/I08-2077

[39] U. K. Sikdar, B. Barik, and B. Gambäck, 'Named Entity Recognition on Code-Switched Data Using Conditional Random Fields', in *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, G. Aguilar, F. AlGhamdi, V. Soto, T. Solorio, M. Diab, and J. Hirschberg, Eds., Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 115–119. doi: 10.18653/v1/W18-3215.

[40] M. Zhang, G. Geng, and J. Chen, 'Semi-Supervised Bidirectional Long Short-Term Memory and Conditional Random Fields Model for Named-Entity Recognition Using Embeddings from Language Models Representations', *Entropy*, vol. 22, p. 252, Feb. 2020, doi: 10.3390/e22020252.

[41] Z. Liu *et al.*, 'Entity recognition from clinical texts via recurrent neural network', *BMC Med. Inform. Decis. Mak.*, vol. 17, no. 2, p. 67, Jul. 2017, doi: 10.1186/s12911-017-0468-7.

[42] X. Dong, L. Qian, Y. Guan, L. Huang, Q. Yu, and J. Yang, 'A multiclass classification method based on deep learning for named entity recognition in electronic medical records', in *2016 New York Scientific Data Summit (NYSDS)*, Aug. 2016, pp. 1–10. doi: 10.1109/NYSDS.2016.7747810.

[43] A. Vaswani *et al.*, 'Attention Is All You Need', Aug. 01, 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[44] L. Schmidt, J. Weeds, and J. P. T. Higgins, 'Data Mining in Clinical Trial Text: Transformers for Classification and Question Answering Tasks', Jan. 30, 2020, *arXiv*: arXiv:2001.11268. doi: 10.48550/arXiv.2001.11268.

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[46] C. Martin, H. Yang, and W. Hsu, 'KDDIE at SemEval-2022 Task 11: Using DeBERTa for Named Entity Recognition', in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, Eds., Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1531–1535. doi: 10.18653/v1/2022.semeval-1.210.

[47] J. Li, A. Sun, J. Han, and C. Li, 'A Survey on Deep Learning for Named Entity Recognition', *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Jan. 2022, doi: 10.1109/TKDE.2020.2981314.

[48] Q. Teng, Z. Liu, Y. Song, K. Han, and Y. Lu, 'A survey on the interpretability of deep learning in medical diagnosis', *Multimed. Syst.*, vol. 28, no. 6, pp. 2335–2355, Dec. 2022, doi: 10.1007/s00530-022-00960-4.

[49] B. J. Gutiérrez *et al.*, 'Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again', Nov. 05, 2022, *arXiv*: arXiv:2203.08410. doi: 10.48550/arXiv.2203.08410.

[50] Y. Hu *et al.*, 'Zero-shot Clinical Entity Recognition using ChatGPT', May 15, 2023, *arXiv*: arXiv:2303.16416. doi: 10.48550/arXiv.2303.16416.

[51] R. Bunescu and M. Pasca, 'Using Encyclopedic Knowledge for Named Entity Disambiguation'.

[52] X. Han and J. Zhao, 'Named entity disambiguation by leveraging wikipedia semantic knowledge', in *Proceedings of the 18th ACM conference on Information and knowledge management*, in CIKM '09. New York, NY, USA: Association for Computing Machinery, Nov. 2009, pp. 215–224. doi: 10.1145/1645953.1645983.

[53] S. Cucerzan, 'Large-Scale Named Entity Disambiguation Based on Wikipedia Data', in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 708–716. Accessed: Jun. 22, 2023. [Online]. Available: https://aclanthology.org/D07-1074

[54] D. Milne and I. H. Witten, 'Learning to link with wikipedia', in *Proceedings of the 17th ACM conference on Information and knowledge management*, Napa Valley California USA: ACM, Oct. 2008, pp. 509–518. doi: 10.1145/1458082.1458150.

[55] P. H. Martins, Z. Marinho, and A. F. T. Martins, 'Joint Learning of Named Entity Recognition and Entity Linking', Jul. 18, 2019, *arXiv*: arXiv:1907.08243. Accessed: Jun. 21, 2023. [Online]. Available: http://arxiv.org/abs/1907.08243

[56] S. Tedeschi, S. Conia, F. Cecconi, and R. Navigli, 'Named Entity Recognition for Entity Linking: What Works and What's Next', in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2584–2596. doi: 10.18653/v1/2021.findings-emnlp.220.

[57] A. Alhelbawy and R. Gaizauskas, *Named Entity Disambiguation Using HMMs*. 2013, p. 162. doi: 10.1109/WI-IAT.2013.173.

[58] M. Srinivasan and D. Rafiei, 'Location-Aware Named Entity Disambiguation', in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, in CIKM '21. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 3433–3438. doi: 10.1145/3459637.3482135.

[59] G. Mobasher, L. Mertová, S. Ghosh, O. Krebs, B. Heinlein, and W. Müller, 'Combining dictionary- and rule-based approximate entity linking with tuned BioBERT', Nov. 11, 2021, *bioRxiv*. doi: 10.1101/2021.11.09.467905.

[60] M. A. Simos and C. Makris, 'Computationally Efficient Context-Free Named Entity Disambiguation with Wikipedia', *Information*, vol. 13, no. 8, Art. no. 8, Aug. 2022, doi: 10.3390/info13080367.

[61] M. Čuljak, A. Spitz, R. West, and A. Arora, 'Strong Heuristics for Named Entity Linking', Jul. 06, 2022, *arXiv*: arXiv:2207.02824. doi: 10.48550/arXiv.2207.02824.

[62] K. Labusch and C. Neudecker, 'Named Entity Disambiguation and Linking Historic Newspaper OCR with BERT', in *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, Eds., in CEUR Workshop Proceedings, vol. 2696. CEUR-WS.org, 2020. Accessed: Dec. 15, 2023. [Online]. Available: https://ceur-ws.org/Vol-2696/paper_163.pdf

[63] W. Bouarroudj, Z. Boufaida, and L. Bellatreche, 'WeLink: A Named Entity Disambiguation Approach for a QAS over Knowledge Bases', in *Flexible Query Answering Systems*, A. Cuzzocrea, S. Greco, H. L. Larsen, D. Saccà, T. Andreasen, and H.

Christiansen, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 85–97. doi: 10.1007/978-3-030-27629-4_11.

[64] H. T. Nguyen and T. H. Cao, 'Named Entity Disambiguation: A Hybrid Statistical and Rule-Based Incremental Approach', in *The Semantic Web*, J. Domingue and C. Anutariya, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008, pp. 420–433. doi: 10.1007/978-3-540-89704-0_29.

[65] N. Priya, 'A Name Entity Detection and Relation Extraction from Unstructured Data by N-gram Features', Jan. 2015, Accessed: Dec. 15, 2023. [Online]. Available: https://www.academia.edu/84867074/A_Name_Entity_Detection_and_Relation_Extraction_from_Unstructured_Data_by_N_gram_Features

[66] Q. Zhang, F. Li, F. Wang, and Z. Li, 'Named Entity Disambiguation Leveraging Multi-aspect Information', in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 248–255. doi: 10.1109/ICDMW.2015.35.

[67] X. Zhou, Y. Miao, W. Wang, and J. Qin, 'A Recurrent Model for Collective Entity Linking with Adaptive Features', *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 01, Art. no. 01, Apr. 2020, doi: 10.1609/aaai.v34i01.5367.

[68] T. Bonomo, 'A Deep Learning approach to real-world Entity Linking: extracting and matching organisation mentions from unstructured text', Jan. 2022, Accessed: Dec. 15, 2023. [Online]. Available: https://aaltodoc.aalto.fi/handle/123456789/112889

[69] H. Dong *et al.*, 'Ontology-driven and weakly supervised rare disease identification from clinical notes', *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, p. 86, May 2023, doi: 10.1186/s12911-023-02181-9.

[70] R. Volz, J. Kleb, and W. Mueller, 'Towards ontology-based disambiguation of geographical identifiers'.

[71] W. Shen, J. Wang, and J. Han, 'Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions', *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 443–460, Feb. 2015, doi: 10.1109/TKDE.2014.2327028.

[72] S. R. Eddy, 'What is a hidden Markov model?', *Nat. Biotechnol.*, vol. 22, no. 10, Art. no. 10, Oct. 2004, doi: 10.1038/nbt1004-1315.

[73] H. Li, T. Liu, W.-Y. Ma, T. Sakai, K.-F. Wong, and G. Zhou, Eds., *Information Retrieval Technology: 4th Asia Infomation Retrieval Symposium, AIRS 2008, Harbin, China, January 15-18, 2008 Revised Selected Papers*, vol. 4993. in Lecture Notes in Computer Science, vol. 4993. Berlin, Heidelberg: Springer, 2008. doi: 10.1007/978-3-540-68636-1.

[74] S. Atmakuri, B. Shahi, A. Rao B, and M. N, 'A comparison of features for POS tagging in Kannada', *Int. J. Eng. Technol.*, vol. 7, pp. 2418–2421, Sep. 2018, doi: 10.14419/ijet.v7i4.14900.

[75] S. Balaji and S. Sasikala, 'Signet: Web Information Retrieval with NE Disambiguation based on HMM and CRF', *Int. J. Mach. Learn. Comput.*, pp. 443–445, 2012, doi: 10.7763/IJMLC.2012.V2.163.

[76] A. Alokaili and M. E. B. Menai, 'SVM ensembles for named entity disambiguation', *Computing*, vol. 102, no. 4, pp. 1051–1076, Apr. 2020, doi: 10.1007/s00607-019-00748-x.

[77] M. Badieh Habib Morgan and M. Van Keulen, *A Generic Open World Named Entity Disambiguation Approach for Tweets*. 2013.

[78] J. Liu, R. Xu, Q. Lu, and J. Xu, 'Explore Chinese Encyclopedic Knowledge to Disambiguate Person Names', in *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, Tianjin, China: Association for Computational Linguistics, Dec. 2012, pp. 138–145. Accessed: Dec. 19, 2023. [Online]. Available: https://aclanthology.org/W12-6326

[79] T. Washio, E. Suzuki, K. M. Ting, and A. Inokuchi, Eds., *Advances in Knowledge Discovery and Data Mining: 12th Pacific-Asia Conference, PAKDD 2008 Osaka, Japan, May 20-23, 2008 Proceedings*, vol. 5012. in Lecture Notes in Computer Science, vol. 5012. Berlin, Heidelberg: Springer, 2008. doi: 10.1007/978-3-540-68125-0.

[80] F. Borchert and M.-P. Schapranow, 'HPI-DHC @ BioASQ DisTEMIST: Spanish Biomedical Entity Linking with Pre-trained Transformers and Cross-lingual Candidate Retrieval'.

[81] N. Kolitsas, O.-E. Ganea, and T. Hofmann, 'End-to-End Neural Entity Linking', in *Proceedings of the 22nd Conference on Computational Natural Language Learning*, A. Korhonen and I. Titov, Eds., Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 519–529. doi: 10.18653/v1/K18-1050.

[82] E. Boros *et al.*, 'Robust Named Entity Recognition and Linking on Historical Multilingual Documents', in *Conference and Labs of the Evaluation Forum (CLEF 2020)*, in Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, vol. 2696. Thessaloniki, Greece: CEUR-WS Working Notes, Sep. 2020, pp. 1–17. doi: 10.5281/zenodo.4068074.

[83] R. Navigli and S. P. Ponzetto, 'BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network', *Artif. Intell.*, vol. 193, pp. 217–250, Dec. 2012, doi: 10.1016/j.artint.2012.07.001.

[84] A. Moro, A. Raganato, and R. Navigli, 'Entity Linking meets Word Sense Disambiguation: a Unified Approach', *Trans. Assoc. Comput. Linguist.*, vol. 2, pp. 231–244, Dec. 2014, doi: 10.1162/tacl_a_00179.

[85] A. Moro, F. Cecconi, and R. Navigli, 'Multilingual Word Sense Disambiguation and Entity Linking for Everybody'.

[86] R. Elmasri and S. Navathe, *Fundamentals of database systems*, Seventh edition. Boston Munich: Pearson, 2016.

[87] T. Connolly and C. Begg, 'Database systems. A practical approach to design, implementation and management', 2010.

[88] Z. M. Ma, S. Lu, and F. Fotouhi, 'Conceptual Data Models for Engineering Information Modeling and Formal Transformation of EER and EXPRESS-G', in *Conceptual Modeling - ER 2003*, I.-Y. Song, S. W. Liddle, T.-W. Ling, and P. Scheuermann, Eds., Berlin, Heidelberg: Springer, 2003, pp. 573–575. doi: 10.1007/978-3-540-39648-2_47.

[89] E. F. Codd, 'A Relational Model of Data for Large Shared Data Banks', vol. 13, no. 6, 1970.

[90] A. Bhattacharyya and D. Chakravarty, 'Graph Database: A Survey', in *2020 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, 2020, pp. 1–8. doi: 10.1109/ICCECE48148.2020.9223105.

[91] HaerderTheo and ReuterAndreas, 'Principles of transaction-oriented database recovery', *ACM Comput. Surv. CSUR*, Dec. 1983, doi: 10.1145/289.291.

[92] M. Demba, 'Algorithm for Relational Database Normalization Up to 3NF', *Int. J. Database Manag. Syst.*, vol. 5, no. 3, pp. 39–51, Jun. 2013, doi: 10.5121/ijdms.2013.5303.

[93] W. H. Allen, 'The Rise of the Botanical Database', *BioScience*, vol. 43, no. 5, pp. 274–279, 1993, doi: 10.2307/1312059.

[94] E. M. Manhã, M. C. Silva, M. G. C. Alves, M. B. Almeida, and M. G. L. Brandão, 'PLANT: A bibliographic database about medicinal plants', *Rev. Bras. Farmacogn.*, vol. 18, pp. 614–617, Dec. 2008, doi: 10.1590/S0102-695X2008000400020.

[95] A. H. Syed and T. Khan, 'SHPIS: A Database of Medicinal Plants from Saudi Arabia', *Int. J. Adv. Comput. Sci. Appl. Ijacsa*, vol. 8, no. 5, Art. no. 5, 48/31 2017, doi: 10.14569/IJACSA.2017.080507.

[96] R. Angles and C. Gutierrez, 'Survey of graph database models', *ACM Comput. Surv.*, vol. 40, Feb. 2008, doi: 10.1145/1322432.1322433.

[97] E. Spadini, F. Tomasi, and G. Vogeler, Eds., *Graph data-models and semantic web technologies in scholarly digital editing*. in Schriften des Instituts für Dokumentologie und Editorik, no. Band 15. Norderstedt: BoD – Books on Demand, 2021.

[98] K. Singh *et al.*, 'DISPEL: database for ascertaining the best medicinal plants to cure human diseases', *Database J. Biol. Databases Curation*, vol. 2023, Oct. 2023, doi: 10.1093/database/baad073.

[99] F. Chelazzi and S. Bonzano, *Thinking data. Integrative big data approaches towards an 'introspective' digital archaeology in the ancient Mediterranean*. EUT Edizioni Università di Trieste, 2020. Accessed: Jun. 10, 2024. [Online]. Available: http://hdl.handle.net/10077/30227

[100] F. Abad-Navarro, J. A. Bernabé-Diaz, A. García-Castro, and J. T. Fernandez-Breis, 'Semantic Publication of Agricultural Scientific Literature Using Property Graphs', *Appl. Sci.*, vol. 10, no. 3, Art. no. 3, Jan. 2020, doi: 10.3390/app10030861.

[101] T. R. Gruber, 'Towards Principles for the Design of Ontologies Used for Knowledge Sharing', in *Formal Ontology in Conceptual Analysis and Knowledge Representation*, N. Guarino and R. Poli, Eds., Deventer, The Netherlands: Kluwer Academic Publishers, 1993.

[102] S. H. Kassani and P. H. Kassani, 'Building an Ontology for the Domain of Plant Science using Prot\'eg\'e', Oct. 11, 2018, *arXiv*: arXiv:1810.04606. Accessed: Dec. 06, 2022. [Online]. Available: http://arxiv.org/abs/1810.04606

[103] R. L. Walls *et al.*, 'The Plant Ontology Facilitates Comparisons of Plant Development Stages Across Species', *Front. Plant Sci.*, vol. 10, 2019, Accessed: Dec. 06, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpls.2019.00631

[104] P. J. Silvie, P. Martin, M. Huchard, P. Keip, A. Gutierrez, and S. Sarter, 'Prototyping a Knowledge-Based System to Identify Botanical Extracts for Plant Health in Sub-Saharan Africa', *Plants*, vol. 10, no. 5, Art. no. 5, May 2021, doi: 10.3390/plants10050896.

[105] C. Carpineto and G. Romano, 'Using Concept Lattices for Text Retrieval and Mining', in *Formal Concept Analysis: Foundations and Applications*, B. Ganter, G. Stumme, and R. Wille, Eds., in Lecture Notes in Computer Science. , Berlin, Heidelberg: Springer, 2005, pp. 161–179. doi: 10.1007/11528784_9.

[106] J. Poelmans, S. Kuznetsov, D. Ignatov, and G. Dedene, 'Review: Formal Concept Analysis in knowledge processing: A survey on models and techniques', *Expert Syst. Appl. Int. J.*, vol. 40, pp. 6601–6623, Nov. 2013, doi: 10.1016/j.eswa.2013.05.007.

[107] A. Braud, X. Dolques, P. Fechter, N. Lachiche, F. Le Ber, and V. Pitchon, 'Analyzing the composition of remedies in ancient pharmacopeias with FCA', in *RealDataFCA'2021, ICFCA Workshop, Strasbourg, France*, in CEUR Workshop Proc. 3151. 2021, pp. 28–35. [Online]. Available: https://ceur-ws.org/Vol-3151/short4.pdf

[108] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*. Springer Verlag, 1999.

[109] P. Allard, S. Ferré, and O. Ridoux, 'Discovering Functional Dependencies and Association Rules by Navigating in a Lattice of OLAP Views', Oct. 2010, pp. 199–210.

[110] A. Braud, C. Nica, C. Grac, and F. Le Ber, 'A lattice-based query system for assessing the quality of hydro-ecosystems', in *CLA 2011*, V. V. A Napoli, Ed., Nancy, France: INRIA NGE et LORIA, Oct. 2011, pp. 265–277. Accessed: May 10, 2022. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00640048

[111] O. Kahl, *Sābūr ibn Sahl's Dispensatory in the Recension of the ʿAḍudī Hospital*. Brill, 2008. Accessed: Jul. 26, 2022. [Online]. Available: https://brill.com/view/title/15688

[112] O. Kahl, *The Dispensatory of Ibn at-Tilmīḏ: Arabic Text, English Translation, Study and Glossaries*. BRILL, 2007.

[113] G. Bos, *Ibn al-Jazzar's Zad al-musafir wa-qut al-hadir, Provision for the Traveller and Nourishment for the Sedentary, Book 7 (7–30) Critical Edition of the Arabic Text with English Translation, and Critical Edition of Moses ibn Tibbon's Hebrew Translation (Ṣedat ha-Derakhim)*. 2015. doi: 10.1163/9789004288614.

[114] E. Loper and S. Bird, 'NLTK: The Natural Language Toolkit', May 17, 2002, *arXiv*: arXiv:cs/0205028. Accessed: Aug. 31, 2024. [Online]. Available: http://arxiv.org/abs/cs/0205028

[115] Y. Liu *et al.*, 'Roberta: A robustly optimized bert pretraining approach', *ArXiv Prepr. ArXiv190711692*, 2019.

[116] A. Conneau *et al.*, 'Unsupervised Cross-lingual Representation Learning at Scale', *CoRR*, vol. abs/1911.02116, 2019, [Online]. Available: http://arxiv.org/abs/1911.02116

[117] T. Wolf *et al.*, 'Transformers: State-of-the-Art Natural Language Processing', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[118] D. Vrandečić, L. Pintscher, and M. Krötzsch, 'Wikidata: The Making Of', in *Companion Proceedings of the ACM Web Conference 2023*, in WWW '23 Companion. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 615–624. doi: 10.1145/3543873.3585579.

[119] R. Agramanisti Azdy and F. Darnis, 'Use of Haversine Formula in Finding Distance Between Temporary Shelter and Waste End Processing Sites', *J. Phys. Conf. Ser.*, vol. 1500, no. 1, p. 012104, Apr. 2020, doi: 10.1088/1742-6596/1500/1/012104.

[120] D. Fernandes and J. Bernardino, 'Graph Databases Comparison: AllegroGraph, ArangoDB, InfiniteGraph, Neo4J, and OrientDB':, in *Proceedings of the 7th International Conference on Data Science, Technology and Applications*, Porto, Portugal: SCITEPRESS - Science and Technology Publications, 2018, pp. 373–380. doi: 10.5220/0006910203730380.

[121] K. El Haff, A. Braud, F. Le Ber, and V. Pitchon, 'Modélisation des ingrédients de remèdes issus de pharmacopées arabes médiévales dans une base de données graphe', *Actes 34es Journ. Francoph. D'Ingénierie Connaiss. Plate-Forme Intell. Artif. PFIA 2023 Jul 2023 Strasbg. Fr.*, p. 34, 2023.

[122] V. Fokou, K. El Haff, A. Braud, X. Dolques, F. Le Ber, and V. Pitchon, 'Exploring Old Arabic Remedies with Formal and Relational Concept Analysis', in *Conceptual Knowledge Structures*, I. P. Cabrera, S. Ferré, and S. Obiedkov, Eds., Cham: Springer Nature Switzerland, 2024, pp. 302–318. doi: 10.1007/978-3-031-67868-4_20.

[123] X. Dolques, A. Braud, M. Huchard, and F. Le Ber, 'RCAExplore, a FCA based Tool to Explore Relational Data', in *Workshop "Applications and Tools of Formal Concept Analysis" (ICFCA)*, Frankfurt, Germany, Jun. 2019. Accessed: Sep. 01, 2024. [Online]. Available: https://hal.science/hal-02446536

[124] A. Braud, X. Dolques, M. Huchard, and F. L. Ber, 'Generalization effect of quantifiers in a classification based on relational concept analysis', *Knowl-Based Syst*, vol. 160, pp. 119–135, 2018.

[125] W. Antoun, F. Baly, and H. Hajj, 'AraBERT: Transformer-based Model for Arabic Language Understanding', in *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, 2020, p. 9.

Appendix A: Maison de la Sagesse, a game inspired by this thesis[21]

(Pour jouer directement sur le web, le mieux c'est Google Chrome et dézoomez la page web si la résolution du jeu est trop grande)

Histoire à savoir avant de jouer :

En 1258, un village d'une terre lointaine est frappé par une maladie terrible, infligeant des souffrances intenses avant de donner la mort aux villageois : visages paralysés... maux de têtes extrêmes... vomissements répétés...

Une villageoise, déterminée à trouver un remède, entreprend un voyage loin de chez elle pour sauver sa famille et son peuple...

Durant son périple, un sage d'une cité étrangère lui parle d'une majestueuse bibliothèque à Bagdad, la Maison de la Sagesse. Dans ce lieu se trouveraient des savoirs remarquables permettant de guérir toutes les maladies.

Arrivée à Bagdad, la villageoise n'y trouve que tristesse et désolation... La ville est en ruine et la bibliothèque brûlée et en partie détruite ! Les manuscrits jetés dans le fleuve du Tigre rendant même les eaux noires d'encre...

Elle apprend alors que ce sinistre tableau est l'œuvre des Mongols, ayant envahi la ville et brisé toute cette civilisation.

Accablée, désespérée, la voyageuse explore les ruines de la bibliothèque, à la recherche de tout indice pouvant la mener à un remède pour son peuple...

En parcourant les fragments de savoirs éparpillés dans les ruines, elle commence à reconstituer les pages de manuscrits qui contiennent les précieuses connaissances. Elle récolte ainsi les fragments avec des dessins des symptômes de son peuple, et entreprend alors de retrouver le remède qu'elle recherche...

Mais il y a un problème : elle ne parle ni ne lit l'arabe.

Réalisé à Strasbourg lors de la Scientific Game Jam 2023, par :

Florian Duta - Game Design

Solène Falk - Music / Sound Design

Karim El Haff - Scientifique, Game Design / Music

Joséphine Herbelin - Art

Antoine Latour - Programmation

Vivien Pfrimmer - Programmation

Antoine Schmoll - Art

 # Main quest completed

[21] The game is accessible online via https://random-sxb.itch.io/la-maison-de-la-sagesse

Université de Strasbourg

Karim EL HAFF

Extraction et modélisation d'informations textuelles pour l'exploitation de pharmacopées arabes anciennes

École doctorale Mathématiques, sciences de l'information et de l'ingénieur | ED 269
Université de Strasbourg

Résumé

Les pharmacopées anciennes, en particulier celles de l'époque abbasside, représentent une source importante de connaissances scientifiques. Cette période fut marquée par une étude approfondie et un développement de la médecine, ayant laissé un ensemble de pratiques qui ont influencé les fondements de la médecine contemporaine. Ces textes offrent un accès privilégié aux savoirs et aux pratiques médicales de l'époque, susceptibles de révéler des synergies pertinentes pour l'élaboration de solutions aux enjeux médicaux actuels. Cependant, l'ampleur et la complexité de ces pharmacopées, d'un point de vue textuel, posent des défis importants pour les chercheurs modernes visant des analyses systématiques et une extraction de données structurée. Ce travail propose une analyse automatique des manuscrits médicaux arabes médiévaux, visant à extraire des informations exploitables pour les études historiques et la pharmacologie moderne. En appliquant des techniques de fouille de texte, incluant la reconnaissance et le liage d'entités nommées, et ensuite la modélisation de l'information, ce travail structure les données extraites, établissant un lien entre ingrédients, remèdes et ressources de données scientifiques contemporaines. Un pipeline de traitement complet a ainsi été développé, aboutissant à un système d'interrogation des textes pharmacopiques anciens afin de découvrir de nouvelles connaissances. Ce travail met en évidence le potentiel des connaissances médicinales historiques pour informer la recherche moderne grâce à des données structurées et analysables.

Mots-clés : Traitement Automatique des Langues, fouille de textes, reconnaissance d'entités nommées, liage d'entités nommées, base de données graphe, pharmacopée

Résumé en anglais

The realm of historical manuscripts, particularly those originating from the Abbasid era, offers a treasure trove of knowledge. It was a time when medicine and other sciences were expansively studied and developed, leaving behind a legacy of medical practices that have shaped much of contemporary medicine. These manuscripts provide a window into the medical wisdom and practices of the time that may help us uncover interesting synergies and patterns useful for finding solutions for modern medical problems. However, the vastness and complexity of these translated pharmacopeias present a challenge for modern researchers aiming for systematic analyses and data extraction. This work undertakes an automatic analysis of medieval Arabic medical manuscripts to extract valuable insights for both historical studies and contemporary pharmacology. Utilizing text-mining techniques, including Named Entity Recognition and Linking and proceeding to perform Information Modeling, this work aims to process unstructured historical texts into formally structured data, linking ingredients and remedies to current scientific data resources. A comprehensive processing pipeline was thus developed culminating in a system for querying ancient pharmacopeic texts to uncover new insights. This work highlights the potential of historical medicinal knowledge to inform modern research through structured, analyzable data.

Keywords: Natural Language Processing, Text Mining, Named Entity Recognition, Named Entity Linking, Graph Database, Pharmacopeia