ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240760

Building and Assessing a Named Entity Recognition Resource for Ancient Pharmacopeias

Karim El Haff[a,c], Wissam Antoun[b], Agnès Braud[a], Florence Le Ber[a] and Véronique Pitchon[c]

[a]Université de Strasbourg, ENGEES, CNRS, ICube UMR 7357, F 67000 Strasbourg, France
[b]Inria-Paris, 75012 Paris, France
[c]Université de Strasbourg, CNRS, Archimède UMR 7044, F 67000 Strasbourg
ORCID (Karim El Haff): https://orcid.org/0009-0000-0519-6418, ORCID (Wissam Antoun): https://orcid.org/0000-0001-8021-5834, ORCID (Agnès Braud): https://orcid.org/0000-0003-3614-9141, ORCID (Florence Le Ber): https://orcid.org/0000-0002-2415-7606

Abstract. This research revolves around utilising Named Entity Recognition (NER) to analyse and categorise data from English translations of pharmacopeias from the Abbasid era, noted for its valuable contributions to science and medicine. The main goal of this work, along with publishing this resource freely, is to assess cross-manuscript NER performance by evaluating the NER model's performance on unseen corpora and translation styles, as well as demonstrating the transferability of the NER task on such corpora. Two distinct experiments were conducted, focusing on F1-scores differences from mixing source translators and varying training dataset sizes. In experiments involving mixing translator styles, training on a mix of all available styles while accounting for dataset size yielded the best F1-scores compared to even training on the same style as the testing data, while experiments with dataset sizes show diminishing returns of scaling training datasets compared to varying translation styles. This work attempts to enhance the exploration of the medical knowledge embodied in these texts to facilitate their analysis for knowledge extraction relevant to modern medical practices. Furthermore, this research demonstrates strategies to optimise NER results in this context, forming a juncture between digitising historical information and enabling further explorations in pharmacopeia-related Natural Language Processing research.

1 Introduction

The realm of historical manuscripts, particularly those originating from the Abbasid era, offers a treasure trove of knowledge [9]. 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, primarily written in Arabic and later translated into various languages, such as English, the language of translation we work on, 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 [17]. However, the vastness and complexity of these translated pharmacopeias present a challenge for modern researchers aiming for systematic analyses and data extraction.

The application of Named Entity Recognition (NER) to such texts can act as a catalyst, streamlining the extraction process and making the data more accessible for research, especially in the context of insight extraction [1]. Figure 1 showcases where the NER task lies within the ultimate goal of insight extraction in our efforts of text mining Abbasid corpora. NER aims to classify named entities into predefined categories pertaining to the intended research. In the context of our study, the challenge is to extract and categorise entities related to medical ingredients, symptoms, organs, and preparation types.

Creating a comprehensive and accurate historical corpus destined for the NER task, especially when dealing with English translations of ancient Arabic texts, presents its own set of unique linguistic and methodological challenges. Here are the paramount questions that we reflected upon and answered prior to the creation of the resource:

Textual Variability: Historical documents frequently contain varied linguistic structures that make annotation processes less straightforward. For example, in the following passage, the translation of Ibn al-Jazzar is referencing measles and smallpox using both metonymy (*eruption*) and an anaphora (*affected by **it***). "He should continue to do so until their **eruption** is completed; then he should stop. **Those affected by it** should be given [to drink] barley broth every morning and should be fed with peeled mungo beans [*Phaseolus mungo*] with orache [*Atriplex hortensis*] or gourd or wild-amaranth [*Amaranthus blitum*]" [3]. This raises the issue of choosing which structure should be annotated as a disease symptom. Indeed, a direct symptom entity annotation was the more efficient choice, as the training data is limited.

Archaic and Domain-Specific Vocabulary: Terms and phrases that were commonplace in historical periods might be obsolete today and, thus, less straightforward to standardise. Terms such as *dāniq*, *ratl*, *mithqāl*, *dirham* were used as units of measurement for ingredients and are commonly transliterated by the translator as they would not have an English equivalent. Furthermore, highly specific ingredient names would also commonly be transliterated from Arabic like *kāmakh* or translated into a rough English equivalent, in this case, *vinegar dressing*. These choices raise the issue of inconsistent transliteration orthographies between different translators of different manuscripts (*kāmakh/kāmkh - dāniq/danik*) leading to a potential inconsistency in the training data and adding additional complexities in post-processing linking. As Named Entity Linking (NEL) is a

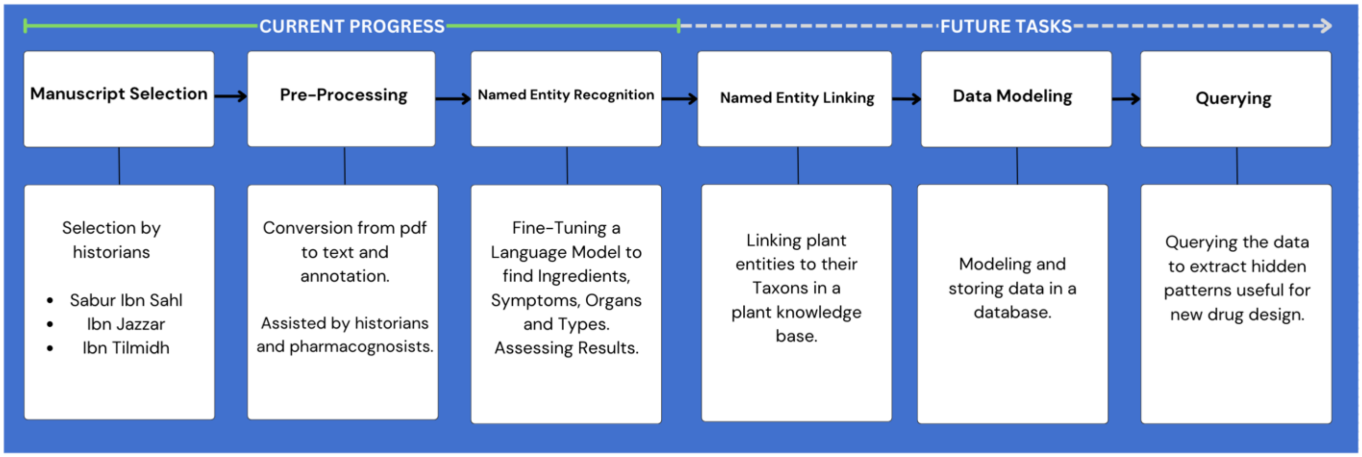

CURRENT PROGRESS
FUTURE TASKS
Manuscript Selection
Pre-Processing
Named Entity Recognition
Named Entity Linking
Data Modeling
Querying
Selection by historians
• Sabur Ibn Sahl
• Ibn Jazzar
• Ibn Tilmidh
Conversion from pdf to text and annotation.
Assisted by historians and pharmacognosists.
Fine-Tuning a Language Model to find Ingredients, Symptoms, Organs and Types. Assessing Results.
Linking plant entities to their Taxons in a plant knowledge base.
Modeling and storing data in a database.
Querying the data to extract hidden patterns useful for new drug design.

Figure 1. The chain of tasks for text mining historical pharmacopeias

complementary task to NER, we plan on tackling it in the future.

Granularity of Annotation: Deciding the level of detail in annotations is an important step as it would have implications on the future usability of the created corpus. While fine-grained annotations can provide richer data ("grilled apple" could have 2 entities: *PREPARATION_METHOD + FRUIT*), they demand more time and expertise, potentially leading to inconsistencies. Coarse-grained annotations ("grilled apple" is an *INGREDIENT*), although quicker and more efficient for training NER models with limited data, might benefit from further steps such as entity linking to capture all the nuances desired for future use cases of the corpus. We chose the latter method, as this work lies within a larger project of text mining where the issue NEL will be tackled in the future.

Quality of Source Materials: The physical quality of historical manuscripts can vary greatly. Faded ink, damaged pages, or poor handwriting can hinder accurate transcription and translation (Figure 2 shows an example folio from the Abbasid era [16]). As some inconsistencies might be present, visually similar words, when written by hand in the Arabic language that uses diacritics to differentiate phonemes, such as *zanbaq (Lily plant)* and *zi'baq (mercury element)* would vastly change the medical and historical implications of the translated remedies, raising the issue of accuracy in corpus creation. Moreover, different copies of the same text might have variations, leading to the challenge of determining authoritative versions. In our case, we decided to work on the English translation of the manuscripts, as the translators are trusted as the choosers of the authoritative versions of the source materials.

Need of Domain Expertise: Historical texts, especially specialised ones like pharmacopeias, demand not just linguistic expertise but also domain knowledge. We collaborated with domain experts, such as historians and botanists to create a reliable corpus, making the process of annotation an interdisciplinary task requiring the effort of multiple people.

Scalability of Manual Annotation: Manual annotation is time-consuming and can be challenging to scale for larger corpora. The complexity of these challenges underscores the importance of a rigorous methodology, collaboration, and iterative refinement when creating and annotating historical corpora for the NER task.

In this work, we harness the power of the transformer-based technology to perform the NER task based on English translations of Abbasid pharmacopeias. As such, our corpus is made up of three manuscripts translated into English from Arabic.

- Dispensatory in the Recension of the 'Aḍudī Hospital" authored by Sābūr ibn Sahl (9th century, Baghdad), translated by Oliver Kahl [13]
- The Dispensatory of Ibn at-Tilmīd (12th century, Baghdad), translated by Oliver Kahl [12]
- Ibn al-Jazzar 's Provision for the Traveller and Nourishment for the Sedentary (10th century, Kairouan), Book 7, translated by Gerrit Bos [3]

Our contribution is thus three-fold:

- Performing a comprehensive investigation into cross-manuscript NER performance by evaluating the model's performance on unseen corpora and translation styles.
- Demonstrating the transferability of the NER task on previously unseen corpora.
- Publishing the model and the training script for future use of fellow researchers.[1]

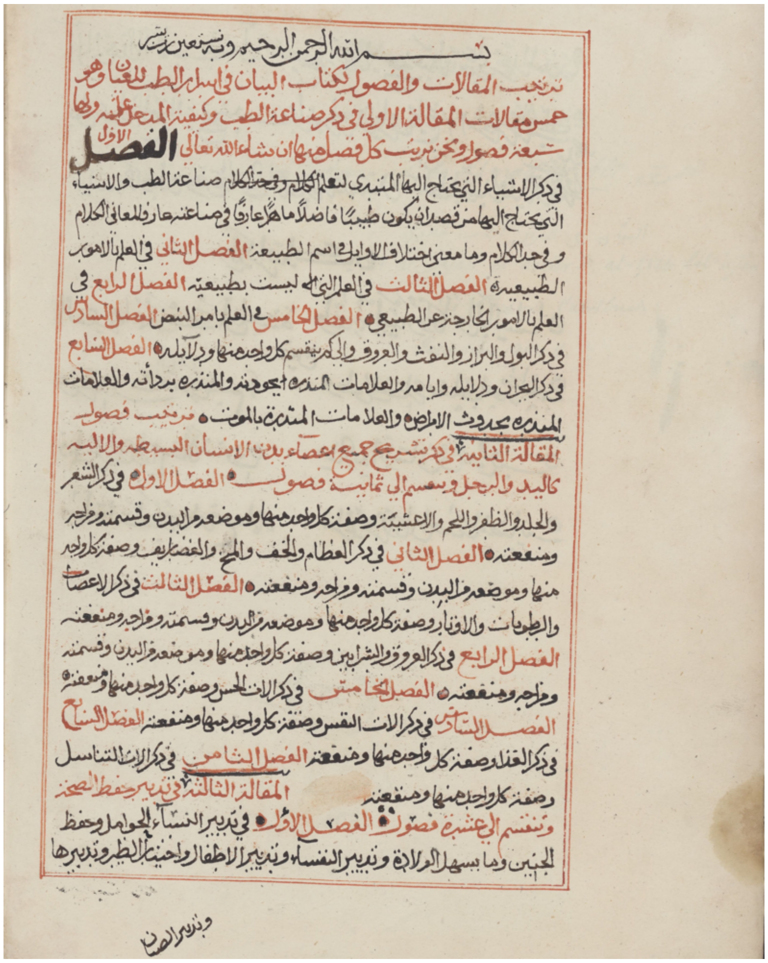
2 Related Work

This section reviews efforts to achieve the NER task in historical documents in general as well as the realm of text mining medical corpora.

2.1 NER in Historical Texts

Historical texts, characterised by their distinct challenges have been the subject of several Named Entity Recognition (NER) studies throughout the years [6]. While the literature is lacking projects tackling the issue of NER for ancient pharmacopeias, to which we would like to contribute, many efforts explore various methodologies, data, and contexts to apply NER to digitised, archival documents.

[1] The resource is fully available on Huggingface via https://huggingface.co/karimelhaff/remed-ner



بسم الله الرحمن الرحيم وبه نستعين رب يسر
ترتيب المقالات والفصول لكتاب البيان في اسرار الطب للعيان وهو
خمس مقالات المقالة الاولى في ذكر صناعة الطب وكيفية المدخل على علمه وفيها
سبعة فصول ونحن نرتب كل فصل منها ان شاء الله تعالى الفصل الاول
في ذكر الاشياء التي يحتاج اليها المبتدي لتعلم ~~الكلام وفي حد الكلام~~ صناعة الطب والاشياء
التي يحتاج اليها من قصد ان يكون طبيبا فاضلا ماهرا عارفا في صناعته عارفا للمعاني الكلام
وفي حد الكلام وما معنى اختلاف الاوايل في اسم الطبيعة الفصل الثاني في العلم بالامور
الطبيعية الفصل الثالث في العلم التي ليست بطبيعية الفصل الرابع في
العلم بالامور الخارجة عن الطبيعي الفصل الخامس في العلم بامر النبض الفصل السادس
في ذكر البول والبراز والنفس والعرق والى كم ينقسم كل واحد منها ودلايله الفصل السابع
في ذكر البحران ودلايله وايامه والعلامات المنذرة بجودته والمنذرة برداته والعلامات
المنذرة بحدوث الامراض والعلامات المنذرة بالموت ترتيب فصول
المقالة الثانية في ذكر تشريح جميع اعضاء بدن الانسان البسيطة والالية
كاليد والرجل وتنقسم الى ثمانية فصول الفصل الاول في ذكر الشعر
والجلد والظفر واللحم والاغشية وصفة كل واحد منها وموضعه من البدن وقسمته ومن اجله
ومنفعته الفصل الثاني في ذكر العظام والغضف والمخ والغضاريف وصفة كل واحد
منها وموضعه من البدن وقسمته ومن اجله ومنفعته الفصل الثالث في ذكر الاعصاب
والرباطات والاوتار وصفة كل واحد منها وموضعه من البدن وقسمته ومن اجله ومنفعته
الفصل الرابع في ذكر العروق والشرايين وصفة كل واحد منها وموضعه من البدن وقسمته
ومن اجله ومنفعته الفصل الخامس في ذكر الات الحس وصفة كل واحد منها ومنفعته
الفصل السادس في ذكر الات النفس وصفة كل واحد منها ومنفعته الفصل السابع
في ذكر الغذاء وصفة كل واحد منها ومنفعته الفصل الثامن في ذكر الات التناسل
وصفة كل واحد منها ومنفعته المقالة الثالثة في تدبير حفظ الصحة
وتنقسم الى عشرة فصول الفصل الاول في تدبير النساء الحوامل وحفظ
الجنين وما يسهل الولادة وتدبير النفساء وتدبير الاطفال واختيار الظير وتدبيرها

وتدبير الصبيان

Figure 2. National University Library of Strasbourg, MS 4187, fol. 10 - Strasbourg, France

Firstly, Grover et al. [8] developed a prototype system focusing on the recognition of person and place names in the digitised records of British parliamentary proceedings from the late 17th and early 19th centuries. The system accepts OCR engine output as input and employs a rule-based NER system implemented using LT-XML2 and LT-TTT2 text processing tools, addressing particular challenges and errors associated with OCR data.

Secondly, Ruokolainen and Kettunen [18] undertook a focused exploration into the application of NER for discerning location and person entities within a corpus of Finnish historical newspapers and journals dating from 1771–1929. Utilising Stanford NER, a trainable statistical NER engine, and confronting the intricacies of the Finnish language and varied OCR quality, their experiment leveraged a 500,000-word ground truth sample from the collection.

Furthermore, the work of Karsvall and Borin [14] incorporated place names from medieval charters with geographical data. Due to the significant volume of the material in the Swedish National Archives, manual markup was deemed infeasible. An off-the-shelf NER system, designed for modern Swedish, was used as a methodology to experiment with the automatic extraction of place names from these medieval documents, thus shedding the light on the need for the creation of training data based on historical corpora to yield better results.

Moreover, Won et al. [21] addressed the identification and extraction of toponyms and spatial information within historical text collections. This investigates challenges inherent in historical corpora, such as language changes, spelling variations, and OCR errors, through an evaluation of five NER systems, and introduces a voting system that combined their outputs. The ensemble approach achieved consistent measures of precision and recall across two collections of historical correspondence, indicating potential viability as a methodology for place name recognition in historical corpora.

Further, Yousef et al. [22] tackled NER in the domain of ancient languages, specifically ancient Greek. Faced with obstacles including a lack of annotated data and appropriate infrastructural support for languages with limited modern usage, they trained two automatic NER models utilising transformer-based models, where both models achieved over 82% F1-score on all classes, indicating the viability of transformer-based technology for scarce data.

Collectively, these works underscore the diverse and yet consistently challenging landscape of applying NER to historical documents, as well as the trend of finding entities such as place and person names in the context of historical studies.

2.2 Text Mining Modern and Non-Modern Medical Texts

Text mining, especially in the context of the NER task, has seen notable applications in varied domains, providing a valuable mechanism to identify and utilise structured information from unstructured text. This traversal through modern and ancient medical text mining aims to underline the challenges and opportunities it presents, thereby situating the present research within a broader context.

Beginning with modern medicine, applications of NER have proven important in discerning relevant entities in medical texts, significantly aiding in informational extraction and consequent data analysis. For instance, Lei et al. [15] explore the extraction of four types of entities—clinical problems, procedures, laboratory tests, and medications—from Chinese clinical texts, illustrating the potential of utilising features like word segmentation and section information to achieve notable results, with F-measures reaching up to 93.51%. Concurrently, Yu et al. [23] employ the BioBERT model for NER in Electronic Medical Records, achieving a notable F1-score of 87.10% and showcasing the efficacy of embedding models in mining entities from medical texts. These examples underscore how NER has become integral in extracting useful information from modern medical documents.

While NER's applications in modern contexts have been expansive, its utilisation in ancient medicine text mining is relatively scarce, thus forming a gap this work aims to address. Applications in the domain of ancient medicine predominantly veer towards generalised text mining, wherein biological and historical patterns are unearthed without explicit automated entity recognition. For example, Connelly et al. [4] employ network analysis techniques to a 15th-century pharmacopeia that was transformed into a database, revealing patterns in ingredient choice that could potentially reflect biological activity against infectious agents. The KNOMANA project by Silvie et al. [19] uses a similar approach by developing a software prototype to navigate through botanical data with the purpose of replacing synthetic pesticides and antimicrobials with plant-based extracts. Likewise, Zhang et al. [24] perform an expansive literature review and expert consultations to identify and validate classical terms in Traditional Chinese Medicine to identify potential herbs and formulae for diabetic nephropathy from classical medical literature, underscoring a methodology where ancient terms are used as proxies for modern medical conditions.

The world of ancient medicine provides a plethora of unique terms, descriptions, and entities, making it a suitable ground for NER exploration, although scarce in the literature. The work of El Haff et al. [7] presented an approach towards named entity recognition within historical Arabic pharmacopeias, notably focusing on a single medieval medical manuscript – " Dispensatory in the Recension of the 'Aḍudī Hospital " authored by Sābūr ibn Sahl in the 9th century, translated to English by Oliver Kahl [13]. This corpus is a subset of the larger corpus that we present in our paper. The authors experimented with various transformer-based models for entity recognition, achieving the highest F1-score of 86.03% with DeBERTaV3 [11], underlining the feasibility of NER on historical pharmacopeias.

3 Data

Training a robust NER model mandates domain-specific data, which acts as the foundational bedrock for the model's comprehension. In this context, we decided to annotate the entirety of remedies from 3 pharmacopeias: Ibn al-Jazzar 's Provision for the Traveller and Nourishment for the Sedentary, Book 7, translated by Gerrit Bos, as well as the Dispensatory in the Recension of the 'Aḍudī Hospital " authored by Sābūr ibn Sahl and The Dispensatory of Ibn at-Tilmīd, both translated by Oliver Kahl. Each translator adopted a different editing style while translating the original pharmacopeias. Oliver Kahl adopted a more structured approach when describing the remedies where the recipe title is presented followed by the symptoms of a disease it treats then the ingredients it contains along with further clarification words added between brackets within the text, with a tendency of leaving some discussion elements as footnotes that we did not take into account in the annotation process to avoid noise. Gerrit Bos' text, on the other hand, 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.

Editing old manuscripts is indeed an intricate 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 3 showcases the different translation styles. Transitioning from a physical or digital manuscript to a usable corpus for Natural Language Processing (NLP) poses challenges. Our starting point was the PDF version of the translated work. To make it amenable for computational processes, we employed a mainstream PDF-to-text tool, facilitating its conversion to a plain text format. With the corpus in a more pliable form, we manually discarded unneeded elements, notably footnotes, introductions, prefaces and every section outside the scope of a pharmacopeia. We then performed tokenization using the NLTK library [2] to make it annotation-ready.

To ensure the quality and reliability of annotations, the corpus underwent meticulous manual annotation by one annotator, a Computational Linguist who was supported and advised by the expertise of Pharmacognosist and a Historian. This collaboration spanned 2 months of part-time annotation work and ensured computational relevance, scientific and historical accuracy. Upon completion, we were equipped with a comprehensive corpus of 93K tokens, each associated with a contextual label.

The specifics of these annotations are detailed in Table 1, which showcases a diverse range of entities encapsulated within the corpus.

Our resulting corpus is 2.5 times larger than the one presented in El Haff et al. [7], which we followed for the annotation process and employed 4 label types:

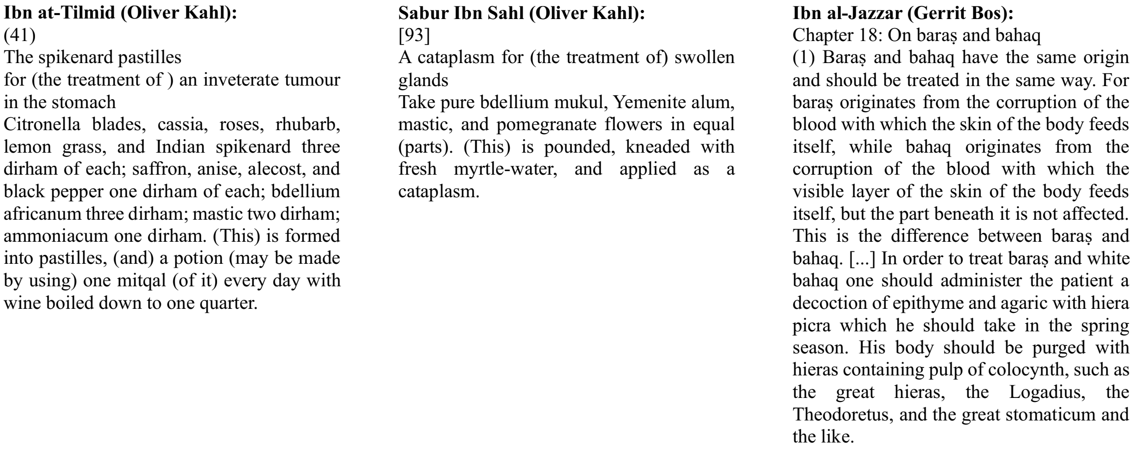
- Type : B-Type I-Type, denoting the remedy's form (e.g., tablet, pill);
- Sym : B-Sym I-Sym, indicating a disease symptom;
- Ing : B-Ing I-Ing, indicating a particular ingredient;
- Org : B-Org I-Org, indicating the mention of an organ or body part;
- O : Tokens that reside outside the purview of the domain-specific entities.

Our choice of these tags is based on the fact that this 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 (INGREDIENT) occur frequently together to treat fevers (SYMPTOM)", "what forms (TYPE) of remedies are most commonly used to treat the skin (ORGAN)?"

To ensure compatibility with prevalent NLP techniques, the data was labeled in the IOB2 format (" inside, outside, beginning "), a prevalent format for token tagging in NLP tasks. The B- prefix demarcates the onset of an entity, the I- prefix continues the entity, and the O label designates tokens that lie outside any specified entity.

Table 1. Tag counts

Dataset	O	ING	SYM	ORG	TYPE	TOTAL TOKENS
Ibn al-Jazzar	12,242	2,874	903	113	146	16,278
Ibn at-Tilmīd	29,983	8,052	1,365	159	771	40,330
Sābūr ibn Sahl	29,094	5,789	1,504	177	396	36,960



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 3. Different translation styles of the corpus' translators

4 Methodology and Experimental Setup

Our NER system is based on the state-of-the-art approach of using a transformer encoder [20, 5] model to classify each input token and assign them to a set of IOB2 labels. Our objective here is to evaluate how well the NER model performance transfers to new unseen manuscripts, given limited annotations. In order to test the model's transfer capabilities using the limited set of annotated manuscripts from section 3, we design our experiments to test the following:

- We test transferability to a new translator, by training our model on the manuscripts translated by Oliver Kahl, namely Ibn at-Tilmīd and Sābūr ibn Sahl, and then testing on the manuscript of Ibn al-Jazzar translated by Gerrit Bos, and vice-versa.
- We look into the effect of transferability between the original authors.
- We study the effect of mixing the training manuscript translators.
- We also examine the effect of the training dataset size.

4.1 Experimental Design

For the goal of testing transferability to a new translator, we have categorised our experiments into three distinct groups to systematically evaluate the performance of our NER model in three different scenarios:

Single Source Manuscript Training. In this set of experiments, we trained our NER model using only one source manuscript, either Ibn al-Jazzar's, Sābūr ibn Sahl's, or Ibn at-Tilmīd's. The purpose was to evaluate the model's performance when trained on a single manuscript and tested on the other two. This case represents the worst-case scenario when the annotations are limited to a single source or translation style.

Combining Two Manuscripts for Training. In these experiments, we combined two source manuscripts for training. This allowed us to assess how well the model generalised when trained on a mixture of sources. We experimented with different combinations, namely Ibn al-Jazzar + Sābūr ibn Sahl, Ibn al-Jazzar + Ibn at-Tilmīd, and Sābūr ibn Sahl + Ibn at-Tilmīd.

Combining All Manuscripts for Training. In this experiment, we combine annotations from all three manuscripts for training and evaluate the model on the test sets from each of the manuscripts individually. This scenario represents a more comprehensive training approach, as the model has access to annotations from all available sources.

Evaluation 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 since we need to account for the variability in the training corpus size, we limit the size of the resulting training corpus to match the smallest manuscript, Ibn al-Jazzar. When combining Sābūr ibn Sahl and Ibn at-Tilmīd for the second experiment, the resulting training corpus consists of a split of the two sources, with the total training example count being equal to the whole Ibn al-Jazzar training set to ensure evaluation consistency.

To examine the effect of the training data size, we repeat the experiments but with the full training corpus used when combining manuscripts. We then evaluated the model's performance on the same test sets containing annotations from all the different manuscripts.

To recap, our experiments aimed to assess the model's ability to generalise across different source manuscripts and to understand how mixing training data from various manuscripts and sizes affected its performance.

4.2 Training Setup

Following El Haff et al. [7], we use the DeBERTaV3 [10] pretrained English monolingual model, as it was the best performer. All models were trained with a batch size of 8, a maximum sequence length of 384 and a linearly decreasing learning rate of 5e-5 for 8 epochs. The

Table 2. F1-scores for experiments with the equal training dataset size (5-seed avg.)

Training set/Text set	**Ibn al-Jazzar**	**Ibn at-Tilmīd**	**Sābūr ibn Sahl**
	(Gerrit Bos)	*(Oliver Kahl)*	*(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

experiments were conducted using five different random seeds, as this approach lowers the potential impact of seed-specific variations on the results.

5 Results

In our pursuit to establish a model exhibiting proficiency in extracting named entities from Abbasid-era pharmacopeias translated into English, our experimental results shown in Table 2 fully addressed the outlined methodological goals.

5.1 Transferability Between Different Translators

The impact of translator variance on model performance is tangible, albeit not drastically so, hinting towards a commendable generalisability of the NER model. By training on manuscripts translated by Oliver Kahl (Ibn at-Tilmīd and Sābūr ibn Sahl) and testing on Gerrit Bos's translation (Ibn al-Jazzar), F1-scores of 69.09 (Ibn at-Tilmīd) and 69.57 (Sābūr ibn Sahl) were achieved, demonstrating a notable yet not debilitating dip when compared to the intra-translator testing score of 75.63 (Ibn al-Jazzar trained and tested). On the other hand, training on Ibn al-Jazzar showed slightly better inter-translator performances (78.25 and 80.55 vs 75.63).

This effect may be due to the nature of Ibn al-Jazzar being less direct and including more discussions compared to the more succinct and structured style of Ibn at-Tilmīd and Sābūr ibn Sahl, as mentioned in Section 3.

5.2 Effect of Original Manuscript Author

Looking at Table 2, training only on manuscripts by Ibn at-Tilmīd or Sābūr ibn Sahl yields similar scores on their respective test sets, suggesting that the model's performance is not affected by the original manuscript author, unlike when changing translators.

5.3 Effect of Mixing Training Manuscripts

Mixing data from different manuscripts clearly offsets the translator effect and slightly boosts the performance overall. Training on a combination of Ibn al-Jazzar and Ibn at-Tilmīd or Ibn al-Jazzar and Sābūr 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 example, combining Ibn al-Jazzar and Ibn at-Tilmīd for training shows an F1-score of 83.58 on the Ibn at-Tilmīd test set, indicating an improvement as compared to training solely on Ibn at-Tilmīd (82.34), or on manuscripts from Oliver Kahl only, namely the Ibn at-Tilmīd and Sābūr ibn Sahl (81.70). The positive effect is also seen when we equally mix all manuscripts, achieving 83.60 F1-score.

5.4 Effect of Training dataset size

Table 3 presents the outcomes of experiments involving training on complete annotated manuscripts. Observing the F1-scores, a pattern emerges: as we utilise more data for training (employing all manuscripts), the model's performance, in general, shows an improvement across different test manuscripts. However, a closer examination reveals an interesting nuance: 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 significantly boost the NER model's performance. This phenomenon reflects a diminishing return on investment when consistently escalating the amount of training data.

5.5 Results by Tag

Table 4 details the scores obtained for each tag for the experiment that is based on full-size manuscripts. Ingredient entities have the best score (0.92), and highest support (1675). Organ entities have a noticeably lower performance compared to the other tags (0.69), possibly due to being the least represented entity in the support (62). For instance, in the sentence "He was stripped from his [diseased] skin when he drank from that wine." The word "wine" was correctly identified as an ingredient while skin was not detected as an organ.

Additionally, Type entities (0.85) perform better than Symptom entities (0.80) even though they have lower support (249 and 335 respectively). This is possibly 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". In the phrase from Ibn Jazzar's manuscript, "Both Galen and Dioscorides have mentioned remedies that should be applied when someone is afraid [of being poisoned]", the entity "being poisoned" that we considered as a symptom was not detected.

Table 3. F1-scores for experiments with the full manuscripts (5-seed avg.)

Training set/Text set	**Ibn al-Jazzar**	**Ibn at-Tilmīd**	**Sābūr ibn Sahl**
	(Gerrit Bos)	*(Oliver Kahl)*	*(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

Tag	Precision	Recall	F1	Support
Ingredient	0.91	0.92	0.92	1675
Organ	0.64	0.74	0.69	62
Symptom	0.77	0.83	0.80	335
Type	0.87	0.84	0.85	249

Table 4. Detailed scores for each tag for the model trained on all manuscripts.

5.6 Discussion

This research sheds light on strategies and hurdles in applying NER models to historical pharmacopeias, particularly under resource constraints. A keen investigation into the manuscripts reveals significant advantages of employing NER models for drawing out and digitally archiving historical information. An important finding of this work underscores the decision-making process pertaining to data acquisition and optimising annotation, as annotating manuscripts is a slow and costly task. When comparing the strategies of augmenting data from a single source and incorporating translations from various authors or translators, our results favour the latter. This approach not only broadens the model's exposure to diverse linguistic expressions and styles but also strengthens its ability to recognise entities across different textual contexts.

Furthermore, our study contributes to the initiatives of digital preservation of ancient pharmacological knowledge and surfaces considerations about the factors that influence the performance of NER models in historical text contexts. This can guide future research in the area of natural language processing, especially concerning historical document analysis and scenarios where resources are scarce.

It is important to recognise that while our results provide clear insights and directions for further research, they also introduce new questions for future studies, particularly in the overlapping field of historical text analysis and NLP. For instance, questions may arise regarding the application of NER models to corpora written in differing time periods, as our corpus is fully based on 21st-century translations of old texts regardless of their age, or exploring how the findings may be impacted by testing on historical medical corpora that were not composed in the form of a pharmacopeia, or documents that were translated from the Greek or Latin medical heritage.

6 Conclusion

In concluding our task of analysing historical pharmacopeias using NER, we underscore the multifaceted implications of our findings, particularly in the broader context of text mining and extracting insightful knowledge from such rich texts.

The purpose of this work lies in the study of the generalisability of transformer-based NER models to different domain-specific authors and translators, as well as the creation of a NER resource for the less-resourced pharmacopeia niche and we believe that its impact lies in the wider domain of information extraction for interested scientists and historians looking to extract insights from old source materials (for new drug development based on ancient knowledge, or understanding ancient drug-making strategies).

In our work, two distinct experiments were conducted, focusing on F1-scores from varying training dataset sizes and full manuscripts. In experiments involving equal training dataset sizes, training on all manuscripts yielded F1-scores of 75.91 ± 0.78, 83.60 ± 0.61, and 83.21 ± 0.97 for different textual sources. When considering experiments with full manuscripts, promising results were observed with F1-scores peaking at 80.91 ± 0.85, 86.51 ± 0.69, and 87.21 ± 0.62 upon training on all manuscripts. The results demonstrate the transferability of the NER task to different translation styles and authors.

We conclude that the best approach to optimise the NER results is through diversification of the training data not only in pharmacopeic content but also in text structure. We also show that scaling training data and annotations has diminishing returns compared to diversification.

Broadly, transforming historical pharmacological knowledge into a digital format has significant implications in data science and natural language processing, especially concerning extracting insights that can inform and enrich our current understanding of ancient medical practices. Text mining, especially from various sources, can reveal hidden patterns, underlying correlations, and potentially rediscover forgotten knowledge or practices that might find relevance in current research and applications.

Through this resource that will be freely published, we want to support initiatives to digitally preserve old yet vital knowledge as well as promoting the exploration of the contents of expansive knowledge in academic archives that potentially contain valuable wisdom.

References

[1] T. Al-Moslmi, M. Gallofré Ocaña, A. Opdahl, and C. Veres. Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access*, 8:32862–32881, Feb. 2020. doi: 10.1109/ACCESS.2020.2973928.

[2] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[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 (edat ha-Derakhim)*. Jan. 2015. ISBN 978-90-04-28847-8. doi: 10.1163/9789004288614.

[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*, 11(1):e03136–19, Feb. 2020. ISSN 2161-2129, 2150-7511. doi: 10.1128/mBio.03136-19. URL https://journals.asm.org/doi/10.1128/mBio.03136-19.

[5] 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)*, pages 4171–4186, 2019.

[6] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, and A. Doucet. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):1–47, Feb. 2021. ISSN 0360-0300, 1557-7341. doi: 10.1145/3604931. URL https://dl.acm.org/doi/10.1145/3604931.

[7] 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. URL https://hal.science/hal-03934557.

[8] C. Grover, S. Givon, R. Tobin, and J. Ball. Named Entity Recognition for Digitised Historical Texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/342_paper.pdf.

[9] R. Hajar. The Air of History Part III: The Golden Age in Arab Islamic Medicine An Introduction. *Heart Views : The Official Journal of the Gulf Heart Association*, 14(1):43–46, 2013. ISSN 1995-705X. doi: 10.4103/1995-705X.107125. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3621228/.

[10] P. He, J. Gao, and W. Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, 2021. _eprint: 2111.09543.

[11] P. He, X. Liu, J. Gao, and W. Chen. DEBERTA: Decoding-Enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

[12] O. Kahl. *The Dispensatory of Ibn at-Tilmī: Arabic Text, English Translation, Study and Glossaries*. BRILL, Mar. 2007. ISBN 978-90-474-1904-4. Google-Books-ID: qNivCQAAQBAJ.

[13] O. Kahl. *Sābūr Ibn Sahl's Dispensatory in the Recension of the Adudi Hospital*. BRILL, 2009. ISBN 978-90-04-17124-4.

[14] O. Karsvall and L. Borin. SDHK meets NER: Linking Place Names with Medieval Charters and Historical Maps. 2018. URL https://www.semanticscholar.org/paper/SDHK-meets-NER%3A-Linking-Place-Names-with-Medieval-Karsvall-Borin/39ae1ac124f4378e0c88d6ad9b83e8417dc1ed6e.

[15] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, and H. Xu. A comprehensive study of named entity recognition in Chinese clinical text. *Journal of the American Medical Informatics Association*, 21(5):808–814, Sept. 2014. ISSN 1067-5027. doi: 10.1136/amiajnl-2013-002381. URL https://doi.org/10.1136/amiajnl-2013-002381.

[16] V. Pitchon and E. Oussiali. Un traité singulier de médecine arabe médiévale :identification d'un manuscrit. *La Revue de la BNU*, 22:51–59, 2020. URL https://hal.science/hal-03085045. (Ms. 4187 de la BNUS, Kitāb al-bayān fī kashf asrār al- ṭibb li-l-iyān d'al-amawī).

[17] V. Pitchon, E. Aubert, C. Vonthron, and P. Fechter. Chapter 14 - how history can help present research of new antimicrobial strategies: the case of cutaneous infections' remedies containing metals from the middle age arabic pharmacopeia. In F. Chassagne, editor, *Medicinal Plants as Anti-Infectives*, pages 459–478. Academic Press, 2022. ISBN 978-0-323-90999-0. doi: https://doi.org/10.1016/B978-0-323-90999-0.00016-1. URL https://www.sciencedirect.com/science/article/pii/B9780323909990000161.

[18] T. Ruokolainen and K. Kettunen. *À la recherche du nom perdu - Searching for Named Entities with Stanford NER in a Finnish Historical Newspaper and Journal Collection*. Apr. 2018.

[19] 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*, 10(5):896, 2021. ISSN 2223-7747. doi: 10.3390/plants10050896. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8146496/.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, \. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008, 2017.

[21] M. Won, P. Murrieta-Flores, and B. Martins. Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, 5, 2018. ISSN 2297-2668. URL https://www.frontiersin.org/articles/10.3389/fdigh.2018.00002.

[22] T. Yousef, C. Palladino, and S. Jänicke. *Transformer-Based Named Entity Recognition for Ancient Greek*. Nov. 2022. doi: 10.13140/RG.2.2.34846.61761.

[23] X. Yu, W. Hu, S. Lu, X. Sun, and Z. Yuan. BioBERT Based Named Entity Recognition in Electronic Medical Record. In *2019 10th International Conference on Information Technology in Medicine and Education (ITME)*, pages 49–52, Aug. 2019. doi: 10.1109/ITME.2019.00022. URL https://ieeexplore.ieee.org/abstract/document/8965108. ISSN: 2474-3828.

[24] L. Zhang, Y. Li, X. Guo, B. H. May, C. C. L. Xue, L. Yang, and X. Liu. Text Mining of the Classical Medical Literature for Medicines That Show Potential in Diabetic Nephropathy. *Evidence-Based Complementary and Alternative Medicine*, 2014:e189125, Mar. 2014. ISSN 1741-427X. doi: 10.1155/2014/189125. URL https://www.hindawi.com/journals/ecam/2014/189125/. Publisher: Hindawi.