Impact of Translation on Biomedical Information Extraction: Experiment on Real-Life Clinical Notes

Christel Gérardin1, MSc, MD; Yuhan Xiong1,2, MS; Perceval Wajsbürt3, PhD; Fabrice Carrat1,4, MD, PhD; Xavier Tannier2, PhD

1Institut Pierre Louis d'Epidémiologie et de Santé Publique, Sorbonne Université, Institut National de la Santé et de la Recherche Médicale, Paris, France
2Shanghai Jiaotong University, Shanghai, China
3Innovation and Data Unit, Assistance Publique Hôpitaux de Paris, Paris, France
4Department of Public Health, Assistance Publique Hôpitaux de Paris, Hôpital Saint-Antoine, Paris, France
5Sorbonne Université, Institut National de la Santé et de la Recherche Médicale, Université Sorbonne Paris-Nord, Laboratoire d'Informatique Médicale et de Connaissances en e-Santé, Paris, France

Corresponding Author:
Christel Gérardin, MSc, MD
Institut Pierre Louis d'Epidémiologie et de Santé Publique, Sorbonne Université, Institut National de la Santé et de la Recherche Médicale
27 rue de Chaligny
Paris, 75012
France
Phone: 33 678148466
Email: christel.ducroz-gerardin@iplesp.upmc.fr

Abstract

Background: Biomedical natural language processing tasks are best performed with English models, and translation tools have undergone major improvements. On the other hand, building annotated biomedical data sets remains a challenge.

Objective: The aim of our study is to determine whether the use of English tools to extract and normalize French medical concepts based on translations provides comparable performance to that of French models trained on a set of annotated French clinical notes.

Methods: We compared 2 methods: 1 involving French-language models and 1 involving English-language models. For the native French method, the named entity recognition and normalization steps were performed separately. For the translated English method, after the first translation step, we compared a 2-step method and a terminology-oriented method that performs extraction and normalization at the same time. We used French, English, and bilingual annotated data sets to evaluate all stages (named entity recognition, normalization, and translation) of our algorithms.

Results: The native French method outperformed the translated English method, with an overall F1-score of 0.51 (95% CI 0.47-0.55), compared with 0.39 (95% CI 0.34-0.44) and 0.38 (95% CI 0.36-0.40) for the 2 English methods tested.

Conclusions: Despite recent improvements in translation models, there is a significant difference in performance between the 2 approaches in favor of the native French method, which is more effective on French medical texts, even with few annotated documents.

Keywords: concept normalization; named entity recognition; natural language processing; translation; translational tool; biomedical data set; bilingual language model

Introduction

Named entity recognition (NER) and term normalization are important steps in biomedical natural language processing (NLP). NER is used to extract key information from textual medical reports, and normalization consists of matching a specific term to its formal reference in a shared terminology such as the United Medical Language System (UMLS) Metathesaurus [1]. Major improvements have been made recently in these areas, particularly for English, as a huge
amount of data is available in the literature and resources. Modern automatic language processing relies heavily on pretrained language models, which enable efficient semantic representation of texts. The development of algorithms such as transformers [2,3] has led to significant progress in this field.

In Figure 1, the term “mention level” indicates that the analysis is carried out at the level of a word or small group of words: first at the NER stage (in blue) and then during normalization (in green); finally, all mentions with normalized concept unique identifiers (CUIs) are aggregated at the “document level” (orange part). The sets of aggregated CUIs per document predicted by the native French and translated English approaches are then compared to the manually annotated gold standard.

In many languages other than English, efforts remain to be made to obtain such results, notably due to a much smaller quantity of accessible data [4]. In this context, our work explores the relevance of a translation step for the recognition and normalization of medical concepts in French biomedical documents. We compared 2 methods: (1) a native French approach where only annotated documents and resources in French are used and (2) a translation-based approach where documents are translated into English, in order to take advantage of existing tools and resources for this language that would allow the extraction of concepts mentioned in unpublished French texts without new training data (zero-shot), as proposed in van Mulligen et al [5].

We evaluated and discussed the results on several French biomedical corpora, including a new set of 42 annotated hospitalization reports with 4 entity groups. We evaluated the normalization task at the document level, in order to avoid a cross-language alignment step at evaluation time, which would add a potential level of error and thus make the results more difficult to interpret (see word alignment in Gao and Vogel [6] and Vogel et al [7]). This normalization was carried out by mapping all terms to their CUI in the UMLS Metathesaurus [1]. Figure 1 summarizes these various stages, from the raw French text and the translated English text to the aggregation and comparison of CUIs at the document level. Our code is available on GitHub [8].

The various stages of our algorithms rely heavily on transformers language models [2]. These models currently represent the state of the art for many NLP tasks, such as machine translation, NER, classification, and text normalization (also known as entity binding). Once trained, these models can represent any specific language, such as biomedical or legal. The power of these models comes from their neural architecture but also largely depends on the amount of data they are trained on. In the biomedical field, 2 main types of data are available: public articles (eg PubMed) and clinical electronic medical record databases (eg MIMIC-III [9]), and the most powerful models are, for example, BioBERT [10], which has been trained on the whole of PubMed in English, and ClinicalBERT [11], which has been trained on PubMed and MIMIC-III. In French, the variety of models is less extensive, with CamemBERT [12] and FlauBERT [13] for the general domain and no specific model available for the biomedical domain.

In Figure 2, axis 1 (green axis on the left) corresponds to the native French branch with a NER step based on a FastText model trained from scratch on French clinical notes and a CamemBERT model. A multilingual Bidirectional Encoder Representations From Transformers (BERT) model was then used for the normalization step, with 2 models tested: a deep multilingual normalization model [14] and CODER [15] with the full version. Axes 2.1 and 2.2 (the 2 purple axes on the right) correspond to the translated English branches, with a first translation step performed by the OPUS-MT-FR-EN model [16] for both. Axis 2.1 (left) was conducted with decoupled NER and normalization steps: FastText trained from PubMed and MIMIC-III [17] for NER, and deep multilingual normalization [14] or CODER [15]
with the English version for normalization. Axis 2.2 (right) used a single system for the NER and normalization stages: MedCAT [18].

In addition to particularly powerful English-language pretrained models, universal biomedical terminologies (ie, the UMLS Metathesaurus) also contain many more English terms than other languages. For example, the UMLS Metathesaurus [1] contains at least 10 times more English terms than French terms, which may enable rule-based models to perform better in English. As mentioned above, each reference concept in the UMLS Metathesaurus [1] is assigned a CUI, associated with a set of synonyms, possibly in several languages, and a semantic group, such as Disorders, Chemicals & Drugs, Procedure, Anatomy, etc.

In parallel, the performance of machine translation has also improved thanks to the same type of transformer-based language models, and recent years have seen the emergence of high-quality machine translations, such as OPUS-MT developed by Tiedemann et al [16], Google Translate, and others. These 2 observations have led several research teams to add a translation step in order to analyze medical texts, for example, to extract relevant mentions in ultrasound reports [19,20] or in the case of the standardization of medical concepts [14,15,21]. Work in the general (nonmedical) domain has also focused on alignment between named entities in parallel bilingual texts [22,23].

Figure 2. Diagram of different experiments comparing French and English language models without and with intermediate translation steps. CHEM: Chemicals & Drugs; CUI: concept unique identifier; DEVI: Devices; DISO: Disorders; EHR: electronic health record; EN: English; FR: French; FT: fine-tuned; PROC: Procedures; UMLS: United Medical Language System.

Methods

Approaches

Overview

Figure 2 shows the main approaches and models used in our study. We explored 1 “native French approach axis” (axis 1 in Figure 2), based on French linguistic models learned from and applied to French annotated data, and 2 “translated English approach axes” (axes 2.1 and 2.2), based on a translation step and concept extraction tools in English. We compared the performance of all axes with the average of the document-level CUI prediction precisions for all documents.

Native French Approach

Axis 1 consisted of 2 stages: a NER stage and a normalization stage. For the NER stage, we used the nested NER algorithm. Next, a normalization step was performed by 2 different algorithms: a deep multilingual normalization model [14] and CODER [15] with the CODER all version.

Translated-English Approach

First, axes 2.1 and 2.2 consisted of a translation step, performed by the state-of-the-art OPUS-MT-FR-EN [16] or Google Translate algorithm. Second, similar to axis 1, axis 2.1 was based on a NER step and a normalization step. The NER step was performed by the same algorithm but trained on the National NLP Clinical Challenges (N2C2) 2019 data set [24] without manual annotation realignment; for the normalization step, we used the same deep multilingual

Axis 2.2 was based on the MedCAT [18] algorithm, which performs NER and normalization simultaneously. In this case, we compared the native French method with a state-of-the-art, ready-to-use English system, which is not available in French.

Data Sets

For all our experiments, we chose to focus on 4 semantic groups of the UMLS Metathesaurus [1]: Chemical & Drugs ("CHEM"), Devices ("DEVI"), corresponding to medical devices such as pacemakers, catheters, etc; Disorders ("DISO"), corresponding to all signs, symptoms, results (eg, positive or negative results of biological tests), and diseases; and Procedures ("PROC"), corresponding to all diagnostic and therapeutic procedures such as imaging, biological tests, operative procedures, etc, as well as the corresponding number of documents.

Table 1 shows the data sets used for all our experiments and the corresponding number of documents. First, 2 French data sets were used for the final evaluation, as well as for training the axis-1 models. QUAERO is a freely available corpus [25] based on pharmacological notes with 2 subcorpora: MEDLINE (short sentences from PubMed abstracts) and EMEA (drug package inserts). We also annotated a new data set of real-life clinical notes from the Assistance Publique Hôpitaux de Paris data warehouse, described in Section S1 in Multimedia Appendix 1.

Table 1. Overview of all data sets used. When a data set is used for both training and testing, 80% of the data set is used for training and 20% is used for testing. Thus, for the EMEA data set, 30 documents were used for training and 8 for testing, 34 French notes were used for training and 8 for testing, and so on.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Languages and data sets</th>
<th>French</th>
<th>English N2C2a 2019</th>
<th>English and French WMTb 2016</th>
<th>WMT 2019 [28]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Drug notices</td>
<td>38</td>
<td>2514</td>
<td>42</td>
<td>&gt;600,000 sent</td>
</tr>
<tr>
<td></td>
<td>MEDLINE titles</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>French notes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>English notes</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MEDLINE titles</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size (documents), n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use</td>
<td>Train NERc</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Test NER</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Normalization</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Test MedCAT</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Translation (fine-tuning)</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Translation (test)</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

aN2C2: National Natural Language Processing Clinical Challenges.
bWMT: Workshop on Machine Translation.
cNER: named entity recognition.

Second, we used the N2C2 2019 corpus [24] with annotated CUIs, on which we automatically added semantic group information from the UMLS Metathesaurus [1], to train the axis-2.1 system and evaluate the NER and English normalization algorithms. We also used the Mantra data set [26], a multilingual reference corpus for biomedical concept recognition.

Finally, we refined and tested the translation algorithms on the Workshop on Machine Translation biomedical corpora of 2016 [27] and 2019 [28]. A detailed description of the number of respective entities in the data sets can be found in Table S1 in Multimedia Appendix 1.

The annotation methods for the French corpus are detailed in Section S1 and Figure S1 in Multimedia Appendix 1. The distribution of entities for this annotation is detailed in Table S1 in Multimedia Appendix 1.

Translation

We used and compared 2 main algorithms for the translation step: the OPUS-MT-FR-EN model [16], which we tested without and with fine-tuning on the 2 biomedical translation corpora of 2016 and 2019 [27,28], and Google Translate as a comparison model.

NER Algorithm

For this step, we used the algorithm of Wajsbürt [29] described in Gérardin et al [30]. This model is based on the representation of a BERT transformer [3] and calculates the scores of all possible concepts to be predicted in the text. The extracted concepts are delimited by 3 values: start, end, and label. More precisely, the encoding of the text corresponds to the last 4 layers of BERT, FastText integration, and a max-pool Char-CNN [31] representation of the word. The decoding step is then performed by a 3-layer long short-term memory [32] with learning weights [33], similar to
the method in Yu et al [34]. A sigmoid function was added to the vertex. Values (start, end, and label) with a score greater than 0.5 were retained for prediction. The loss function was a binary cross-entropy, and we used the Adam optimizer [35].

In our experiments, for the native French axis (axis 1 in Figure 2), the pretrained embeddings used to train the model were based on a FastText model [36], trained from scratch on 5 gigabytes of clinical text, and a CamemBERT-large model [12] fine-tuned on this same data set. For English axis 2.1, the pretrained models were BioWordVec [17] and ClinicalBERT [11].

**Normalization Algorithms**

**Overview**

This stage of our experiments was essential for comparing a method in native French and one translated into English, and it consisted of matching each mention extracted from the text to its associated CUI in the UMLS Metathesaurus [1]. We compared 3 models for this step, described below: the deep multilingual normalization algorithm developed by Wajsbürt et al [14]; CODER [15]; and the MedCAT [18] model, which performs both NER and normalization.

These 3 models require no training data set other than the UMLS Metathesaurus.

**Deep Multilingual Normalization**

This algorithm by Wajsbürt et al [14] considers the normalization task as a highly multiclass classification problem with cosine similarity and a softmax function as the last layer. The model is based on contextual integration, using the pretrained multilingual BERT model [3], and works in 2 steps. In the first step, the BERT model is fine-tuned and the French UMLS terms and their corresponding English synonyms are learned. Then, in the second step, the BERT model is frozen and the representation of all English-only terms (ie, those present only in English in the UMLS Metathesaurus [1]) is learned. The same training is used for the native French and translated English approaches. This model was trained with the 2021 version of the UMLS Metathesaurus [1], corresponding to the version used for annotating the French corpus. The model was thus trained on over 4 million concepts corresponding to 2 million CUIs.

**CODER**

The CODER algorithm [15] was developed by contrastive learning on the basis of the medical knowledge graph of the UMLS Metathesaurus [1], with concept similarities being calculated from the representation of terms and relations in this knowledge graph. Contrastive learning is used to learn embeddings through multisimilarity loss [37]. The authors have developed 2 versions: a multilingual version based on the multilingual BERT [3] and an English version based on the pretrained BioBERT model [10]. We used the multilingual version for axis 1 (native French approach) and the English version for axis 2.1. Both types of this model (CODER all and CODER en) were trained with the 2020 version of UMLS (publicly available models). CODER all [15] was trained on over 4 million concepts corresponding to 2 million CUIs, and CODER en was trained on over 3 million terms and 2 million CUIs.

For the deep multilingual model and the CODER model, in order to improve performance in terms of accuracy, we chose to add semantic group information (ie, *Chemical & Drugs*, *Devices*, *Disorders*, and *Procedures*) to the model output: that is, from the first $k$ CUIs chosen from a mention, we selected the first from the corresponding group.

The MedCAT algorithm is described in detail in Section S1 in Multimedia Appendix 1.

**Ethical Considerations**

The study and its experimental protocol were approved by the Assistance Publique Hôpitaux de Paris Scientific and Ethical Committee (IRB00011591, decision CSE 20-0093). Patients were informed that their electronic health record information could be reused after an anonymization process, and those who objected to the reuse of their data were excluded. All methods were applied in accordance with the relevant guidelines (*Commission nationale de l'informatique et des libertés* reference methodology MR-004 [38]).

**Results**

The sections below present the performance results for each stage. The N2C2 2019 challenge corpus [24] enabled us to evaluate the performance of our English models on clinical data, and the Biomedical Translation 2016 shared task [27] allowed us to evaluate our translation performance on biomedical data with a BLEU score [39].

**NER Performances**

To be able to compare our approaches in native French and translated English, we used the same NER model, trained and tested on each of the data sets described above. Table 2 shows the corresponding results. Overall $F_1$-scores were similar across data sets: from 0.72 to 0.77.
Table 2. Named entity recognition (NER) performance for each model. For all experiments, we used the same NER algorithm but with different pretrained models. The best performance values are italicized.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F₁-score</td>
</tr>
<tr>
<td></td>
<td>CHEMc</td>
<td>0.80</td>
<td>0.83</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>DEVI</td>
<td>0.42</td>
<td>0.81</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>DISO</td>
<td>0.54</td>
<td>0.63</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>PROCf</td>
<td>0.73</td>
<td>0.78</td>
<td>0.74</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>0.71</td>
<td>0.77</td>
<td>0.74</td>
</tr>
</tbody>
</table>

aFastText* corresponds to a FastText model [36] trained from scratch on our clinical data set.
bN2C2: National Natural Language Processing Clinical Challenges.
cCHEM: Chemical & Drugs.
dDEVI: Devices.
eDISO: Disorders.
fPROC: Procedures.

Normalization Performances

This section presents only the normalization performance based on the gold standard’s entity mentions, without the intermediate steps. The results are summarized in Table 3. The deep multilingual algorithm performed better for all corpora tested, with an improvement in F₁-score from +0.06 to +0.11. By way of comparison, the winning team of the 2019 N2C2 had achieved an accuracy of 0.85 using the N2C2 data set directly to train their algorithm [24]. In our context of comparing algorithms between 2 languages, the normalization algorithms were not trained on data other than the UMLS Metathesaurus. MedCAT’s performance (shown in Table S2 in Multimedia Appendix 1) cannot be directly compared with that of other models, as this method performed both NER and normalization in a single step. However, we note that this algorithm performed as well as axis 2.1 in terms of overall performance, as shown in Table 4.

Table 3. Performance of the normalization step. Model results were calculated from the annotated data sets, focusing on the 4 semantic groups of interest: Chemical & Drugs, Devices, Disorders, and Procedures. The best performance values are italicized.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Data set models</th>
<th>EMEA test</th>
<th>French notes</th>
<th>N2C2b 2019 test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep</td>
<td></td>
<td>0.65</td>
<td>0.57</td>
<td>0.74</td>
</tr>
<tr>
<td>multilingual normalization</td>
<td>CODER all</td>
<td>0.58</td>
<td>0.51</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>CODER en</td>
<td>—</td>
<td>—</td>
<td>0.63</td>
</tr>
</tbody>
</table>

aN2C2: National Natural Language Processing Clinical Challenges.
bNot applicable.

Table 4. Overall performances. The normalization step was performed by the deep multilingual model and the translation was performed by the OPUS-MT-FR-EN FT model. The best performance values are italicized.

<table>
<thead>
<tr>
<th>Methods</th>
<th>EMEA test</th>
<th>French notes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F₁-score (95% CI)</td>
</tr>
<tr>
<td>Axis 1 (French NER*+normalization)</td>
<td>0.63</td>
<td>0.60</td>
<td>0.61 (0.53-0.65)</td>
</tr>
<tr>
<td>Axis 2.1 (Translation+NER+normalization)</td>
<td>0.53</td>
<td>0.40</td>
<td>0.45 (0.38-0.51)</td>
</tr>
<tr>
<td>Axis 2.2 (Translation+MedCAT [18])</td>
<td>0.53</td>
<td>0.46</td>
<td>0.49 (0.38-0.54)</td>
</tr>
</tbody>
</table>

aNER: named entity recognition.

Translation Performances

For both translation models, the respective BLEU scores [39] were calculated on the shared 2016 Biomedical Translation Task [27]. The chosen BLEU algorithm was the weighted geometric mean of the n-gram precisions per sentence.

A fine-tuned version of OPUS-MT-FR-EN [16] was also tested on the 2016 and 2019 Biomedical Translation shared tasks. For fine-tuning, we used the following hyperparameters: a maximum sequence length of 128 (mainly for computational memory reasons), a learning rate of 2 × 10⁻⁵, and a weight decay of 0.01, and we varied the number of epochs up to 15 epochs (the error function curve stops decaying after 10 epochs). The Google Translate model could not be used for our clinical score experiments for reasons of confidentiality.
Table 5 presents the BLEU scores for the 3 models, showing that fine-tuning the OPUS-MT-FR-EN model [16] on biomedical data sets gave the best results, with a BLEU score [39] of 0.51. This was the model used to calculate the overall performance of axes 2.1 and 2.2.

<table>
<thead>
<tr>
<th>Models</th>
<th>WMT® Biomed 2016 test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Translate</td>
<td>0.42</td>
</tr>
<tr>
<td>OPUS-MT-FR-EN</td>
<td>0.31</td>
</tr>
<tr>
<td>OPUS-MT-FR-EN FT&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.51</td>
</tr>
</tbody>
</table>

<sup>a</sup>WMT: Workshop on Machine Translation.
<sup>b</sup>OPUS-MT-FR-EN FT corresponds to the OPUS-MT-FR-EN model [16] fine-tuned on biomedical translated corpus from the WMT Biomedical Translation Tasks in 2016 [27] and 2019 [28].

### Discussion

#### Principal Findings

In this paper, we compared 2 approaches for extracting medical concepts from clinical notes: a French approach based on a French language model and a translated English approach, where we compared 2 state-of-the-art English biomedical language models, after a translation step. The main advantages of our experiment are that it is reproducible and that we were able to analyze the performance of each step of the algorithm: NER, normalization, and translation, and to test several models for each step.

#### The Quality of the Translation Is Not Sufficient

We showed that the native French approach outperformed the 2 translated English approaches, even with a small French training data set. This analysis confirms that, where possible, an annotated data set improves feature extraction. The evaluation of each intermediate step showed that the performance of each module was similar in French and English. We can therefore conclude that it is rather the translation phase itself that is of insufficient quality to allow the use of English as a proxy without a loss of performance. This is confirmed by the translation performance calculations, where the calculated BLEU scores were relatively low, although improved by a fine-tuning step.

In conclusion, although translation is commonly used for entity extraction or term normalization in languages other than English [5,20,41-43], due to the availability of turn-key models that do not require additional annotation by a clinician, we showed that this induces a significant performance loss.

Commercial application programming interface–based translation services could not be used for our task due to data confidentiality issues. However, the OPUS-MT model is considered state of the art, it is adjustable to domain-specific data, and the translation results presented in Table 5 confirm the absence of performance difference between this model and the Google Translate model.

Although our experiments were carried out on a single language, the French-English pair is one of the best performers in recent translation benchmarks [16]. Other languages are unlikely to produce significantly better results.

#### Error Analysis

In these experiments, the overall results may appear low, but the task is still complex, especially because the UMLS Metathesaurus [1] contains many synonyms with different CUIs. To better understand this, we performed an error analysis on the normalization task only, as shown in Table S3 in Multimedia Appendix 1, with a physician’s evaluation, on a sample of 100 errors for both models. We calculated that 24% (24/100) and 39% (39/100) of the terms found by the deep normalization algorithm [14] and CODER [15], respectively, were in fact synonyms but had 2 different UMLS CUIs. This highlights the difficulty of achieving normalization on the UMLS Metathesaurus. The UMLS Metathesaurus indeed groups together numerous terminologies whose mapping between terms is often imperfect, implying that certain synonyms, as shown here, do not have the same CUI, as pointed out by Cimino [44] and Jiménez-Ruiz et al [45]. For example, “cardiac ultrasound” has the CUI of C1655737, whereas “echocardiography” has another CUI of C0013516; similarly, “H/O: thromboembolism” has a CUI of C0455533, whereas “history of thromboembolism” has a CUI of C1997787, and so on.

Moreover, to be more precise, each axis had its own errors: overall, the errors in axis 2 were essentially due to the loss of information in translation. One notable error was literal translation: for example, “dispersed lupus erythematosus” instead of “systemic lupus erythematosus,” or “crepitant” instead of “crackles.” This loss of translation led to more errors in the extraction of named entities.
In addition to the loss of translation information, axis 2.1 was also penalized by the NER step, due to the difference between the training set (N2C2 notes) and the test set (the translated French notes; the aim being to compare the performance of English-language turnkey models with the performance of French-language models from an annotated set). Axis 2.1, for example, omitted the names of certain drugs more often.

Finally, both axes were penalized by abbreviations. These were often badly translated (for example, the abbreviation “MFIU” for “mort foetale in utero,” meaning “intrauterine fetal death,” was not translated), which penalized axis 2. Nevertheless, if they were indeed extracted by NER steps in axis 1, they were not correctly normalized due to the absence of a corresponding CUI in the UMLS Metathesaurus.

Limitations
This work has several limitations. First, the actual French clinical notes contained very few terms in the Devices semantic group, which prevented the NER algorithm from finding them in the test data set. However, this drawback, which penalized the native French approach, still allowed us to draw a conclusion for the results. Furthermore, in this study, we did not take into account attributes of the extracted terms such as negation, hypothetical attribute, or belonging to a person other than the patient for comparison purposes, as the QUAERO [25] and N2C2 2019 [24] data sets did not have this labeled information.

Acknowledgments
The authors would like to thank the Assistance Publique Hôpitaux de Paris (AP-HP) data warehouse, which provided the data and the computing power to carry out this study under good conditions. We wish to thank all the medical colleges, including internal medicine, rheumatology, dermatology, nephrology, pneumology, hepatogastroenterology, hematology, endocrinology, gynecology, infectiology, cardiology, oncology, emergency, and intensive care units, that gave their permission for the use of the clinical data.

Data Availability
The data sets analyzed as part of this study are not accessible to the public due to the confidentiality of data from patient files, even after deidentification. However, access to raw data from the Assistance Publique Hôpitaux de Paris (AP-HP) data warehouse can be granted by following the procedure described on its website [46]: by contacting the ethical and scientific committee at secretariat.cse@aphp.fr. Prior validation of access by the local institutional review committee is required. In the case of non-APHP researchers, a collaboration contract must also be signed.

Authors’ Contributions
CG contributed to conceptualization, data curation, formal analysis, investigation, methodology, software, validation, original drafting, writing—original version, and writing—revision and editing the manuscript. YX contributed to investigation, methodology, software, and validation. PW contributed to investigation, software, and revision of the manuscript. FC contributed to conceptualization, methodology, project administration, supervision, writing—original version, and writing—revision and editing of the manuscript. XT contributed to conceptualization, formal analysis, methodology, writing—original version, and writing—revision and editing of the manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Detailed description of the data sets, an example of the clinical notes annotation, French corpus annotation, MedCAT performances, and error analysis.

[DOCX File (Microsoft Word File), 154 KB-Multimedia Appendix 1]

References


Wajsbürt P. Extraction and Normalization of Simple and Structured Entities in Medical Documents [thesis]. Sorbonne Université; Dec 2021. URL: https://theses.hal.science/THESES-SU/tel-03624928v1 [Accessed 2024-03-15]


Abbreviations

- **BERT**: Bidirectional Encoder Representations From Transformers
- **CUI**: concept unique identifier
- **N2C2**: National Natural Language Processing Clinical Challenges
- **NER**: named entity recognition
- **NLP**: natural language processing
- **UMLS**: United Medical Language System