# Medical concept normalization in French using multilingual terminologies and contextual embeddings

Perceval Wajsbürt, Arnaud Sarfati, Xavier Tannier

Version of Record: https://www.sciencedirect.com/science/article/pii/S1532046421000137
Manuscript_7e6592b0818b23bfcf0ee17dac5704e9

# Medical concept normalization in French using multilingual terminologies and contextual embeddings

Perceval Wajsbürt[1], Arnaud Sarfati[2], Xavier Tannier[1]

[1] Sorbonne Université, Inserm, Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, LIMICS, 75006 Paris, France

[2] AP-HP, DSI-WIND, Paris, France

**Contact information:**

Xavier Tannier

15 rue de l'école de médecine

75006 PARIS

phone: +33 1 44 27 91 13

e-mail: xavier.tannier@sorbonne-universite.fr

keywords: Natural Language Processing, Information Extraction, Medical concept normalization, Multilingual representation.

## ABSTRACT

Introduction: Concept normalization is the task of linking terms from textual medical documents to their concept in terminologies such as the UMLS®. Traditional approaches to this problem depend heavily on the coverage of available resources, which poses a problem for languages other than English.

Objective: We present a system for concept normalization in French. We consider textual mentions already extracted and labeled by a named entity recognition system, and we classify these mentions with a UMLS concept unique identifier. We take advantage of the multilingual nature of available terminologies and embedding models to improve concept normalization in French without translation nor direct supervision.

Materials and Methods: We consider the task as a highly-multiclass classification problem. The terms are encoded with contextualized embeddings and classified *via* cosine similarity and softmax. A first step uses a subset of the terminology to finetune the embeddings and train the model. A second step adds the entire target terminology, and the model is trained further with hard negative selection and softmax sampling.

Results: On two corpora from the Quaero FrenchMed benchmark, we show that our approach can lead to good results even with no labeled data at all; and that it outperforms existing supervised methods with labeled data.

Discussion: Training the system with both French and English terms improves by a large margin the performance of the system on a French benchmark, regardless of the way the embeddings were pretrained (French, English, multilingual). Our distantly supervised method can be applied to any kind of documents or medical domain, as it does not require any concept-labeled documents.

Conclusion: These experiments pave the way for simpler and more effective multilingual approaches to processing medical texts in languages other than English.

## 1. INTRODUCTION

Medical concept recognition and normalization is a classical and essential task of natural language processing (NLP) for biomedical applications [1]. The normalization of terms or phrases present in unstructured texts consists of linking these phrases to the concepts of a knowledge base or a biomedical terminology. This task, also called "entity linking," can be seen as a massively multiclass classification task, since reference terminologies can contain several million concepts. This number of possible classes makes the problem difficult from many points of view. In particular, it makes a fully-supervised learning approach using labeled textual documents impossible since it is impossible to annotate a corpus containing all of the concepts.

For this reason, it is necessary to use the terminology itself, and the lists of terms associated with the concepts, to guide the system [2, 3, 4]. This then raises the problem of the availability of this terminological resource in the target language. In the UMLS metathesaurus [5], often used as the reference terminology, less than 4% of the concepts are associated with at least one term in French, yet one of the most represented languages after English. In the corpus Quaero FrenchMed [6], containing French biomedical texts with mentions labeled with UMLS-2014AB concepts, 27% of the concepts have no synonym in French in this version of the UMLS. Efforts have been made to improve this coverage by manual or automatic translation, or by mapping local terminologies, leading to more complete resources out of the official UMLS [7, 8, 9, 10, 11]. However, the gap is still significant, and this represents a real pitfall for the NLP systems in French, and more generally, in all languages other than English [12].

Recent years have seen the advent in NLP of the use of models pretrained on large amounts of data, on tasks such as language modeling, for an application by transfer learning on different tasks with fewer data. In particular, these models make it possible to represent words or phrases in the vector

space (embeddings) by preserving the similarities and the semantic relationships between these words. They have improved state of the art in many tasks, including in the biomedical domain [13, 14]. Some of these models have been pretrained on multilingual data [15, 16, 17], without any alignment or translation mechanism, and showed performance gains for a wide range of cross-lingual transfer tasks.

In this work, we present a contribution to the task of concept normalization in French. We focus on normalizing mentions that are already extracted and labeled (gold standard spans of text). We classify these mentions with a UMLS concept unique identifier. In the end-to-end task of extracting concepts from free, unlabeled text, obtaining these labeled mentions automatically could be achieved with a named entity recognition system trained on French medical data [54,55,56].

We improve medical concept normalization from gold-standard named entities in French; we overcome the difficulty of the limited terminological resources in this language by taking advantage of English terminologies and the latest multilingual embedding models. We propose a hybrid approach using the multilingual nature of both external knowledge from the UMLS and data-driven, pretrained embedding models. We do not use any manually-enhanced and automatically-translated terminology and do not rely on any task-specific rule. We take advantage of the latest progress in neural network optimization, which allows us to cast the normalization task as a standard, highly-multiclass, classification problem. In contrast, other recent deep learning-based approaches work around the problem by converting the task into a nearest neighbor problem and must insert all the available synonyms in the search space. On two corpora from the Quaero FrenchMed benchmark, we show that our approach can lead to good results even with no labeled data at all (distant learning from the terminology only). It outperforms existing methods, including those using machine translation, with labeled data (distant + supervised learning). We also designed a set of experiments

showing the contribution of multilingual models and resources compared to monolingual.[1]

## 2. BACKGROUND

The clinical normalization problem has been extensively studied in clinical informatics literature. Methods have shifted from rule-based and term indexer systems [18, 19, 20, 4] to machine learning systems [21, 22, 23, 24, 25, 26].

The normalization problem is also known in the general domain as entity linking [27, 28], but differs by the fact that the general domain annotated corpora can leverage larger annotated corpora such as Wikipedia. These make it possible to perform a single supervised training and rely on entity frequencies. However, medical lexicons do not provide context nor accurate medical concept frequencies.

Deep synonym similarity approaches [21, 22, 24, 26, 25, 51] reformulate the normalization task as a nearest neighbor or ranking problem. However, all synonym representations must be stored or recomputed at inference time to perform lookup. Another approach is to subset the list of concepts using an efficient search engine and apply a neural similarity model to achieve a more refined classification amongst the candidates [53]. However, the search engine in such methods can fail for entities that share no lexical pattern with their concept synonyms, especially when normalizing in a language other than English.

We chose to consider the task as a standard classification task amongst concepts, meaning that we only encode concepts of the target terminology, rather than their synonyms, into fixed-length

[1] The code for all experiments described in this paper is available at the following URL: https://github.com/percevalw/deep_multilingual_normalization/tree/master

representations that can be stored and even indexed to accelerate lookup at inference time. This way of framing the problem as a standard classification is similar to [52]. However, in our case, we deal with a considerable number of classes (nearly a million in our experiments). We tackle this difficulty by making use of developments in neural network optimization over the last years, such as Adam [29], BatchNorm [30], or the recent advances on the face identification problem [31, 32, 33]. Indeed, face identification is similar to medical normalization in terms of the number of target classes and the low number of examples per class.

Medical named entity representation learning has closely followed the general domain NLP research by using pretrained models such as Word2Vec [34], Elmo [35], BERT, and its multilingual variant [17]. The latter processes the input text, regardless of its language, to compute contextualized token representations.

The normalization of medical entities in languages other than English has so far relied mainly on the translation of English synonyms into the target language [18, 4], or conversely, the translation of entities into English [36, 23]. These systems use processing pipelines that mix term search with software like MetaMap [50] or Apache Solr, and web-service or local translation [37]. We chose to design and evaluate an auto-sufficient deep neural network classifier with few to no preprocessing of the input named entities.

## 3. MATERIAL

### 3.1 UMLS

The Unified Medical Language System® (UMLS®) is a metathesaurus that unifies concepts from several dozen terminologies in the biomedical domain [5]. Each concept in the UMLS is assigned a Concept Unique Identifier (CUI), a set of terms (or synonyms), possibly in multiple languages, and a semantic type. UMLS semantic types are grouped in 15 semantic groups and each concept is associated with one semantic group, with very few exceptions [40]. For example, “Eicosapentanoic

acid" (concept C0000545) is in the chemical (CHEM) group, while "Accountant" (concept C0000937) is in the living beings (LIVB) group. Table 1 shows statistics on the number of concepts and synonyms in English and French, for the versions 2014AB and 2019AB, both used in this work. French is the 2nd (resp. 5th) most represented language in the 2014 (resp. 2019) version in the metathesaurus, but only 3.5% (resp. 3.6%) of the concepts have terms in French. In this article, we will call "English mirror" the synonyms in English for the concepts that also have synonyms in French. We call "English 5 sources" (EN5) the UMLS concepts that have an English synonym and are either in the five CHV, SNOMEDCT_US, MTH, NCI, or MSH terminologies. We chose these terminologies because they cover 96% of the labels in the Quaero training corpus, without exceeding a million labels.

*Table 1. UMLS statistics. The English mirror is the set of concepts having synonyms in both English and French.*

| Version | Language | #synonyms | #concepts | #synonyms / #concepts |
|---|---|---|---|---|
| 2014AB | English | 5,772,518 | 2,528,878 | 2.28 |
| | English 5 sources | 2,298 600 | 766,548 | 3.00 |
| | French | 179,992 | 88,985 | 2.02 |
| | English mirror | 544,383 | 88,911 | 6,12 |
| 2019AB | English | 9,187,793 | 4,258,236 | 2.16 |
| | English 5 sources | 3,055,453 | 968,467 | 3.15 |
| | French | 374,144 | 154,362 | 2.42 |
| | English mirror | 903,098 | 154,307 | 5,85 |

### 3.2 Quaero FrenchMed corpora

The Quaero FrenchMed corpus [11] consists of two sets of textual documents in French, annotated

with concept CUIs from the 2014AB version of the UMLS:

- Titles of research articles indexed in the MEDLINE database
- Information on marketed drugs from the European Medicines Agency (EMEA)

Unlike other normalization corpora such as NCBI [38] or BC5CDR [39], the annotated concepts were not limited to vocabularies such as SNOMED, MeSH, or OMIM. However, they were limited to 10 of the 15 UMLS semantic groups.

We used two different versions of these corpora in our experiments. The first version, that we call EMEA 2015 and Medline 2015, was used for the CLEF eHealth evaluation lab in 2015, a challenge for named entity recognition and concept normalization. The organizers proposed a training set and a test set for this task. In 2016, a new challenge was organized; the 2015 test set was released as a development set, and a new test set was annotated, leading to a larger corpus containing the previous one.

Table 2 presents general corpus statistics including the number of annotated mentions (i.e., text spans linked to UMLS concepts within the documents), the number of unique mentions, the number of unique concept CUIs, as well as the rate of mentions in each corpus that are linked to a concept with at least one synonym in French in the terminology. Note that very few mentions are annotated with more than one CUI in the corpora.

To ensure a fair comparison with other systems published on this benchmark, we use the 2014AB version of the UMLS, unless specified otherwise.

*Table 2. Overview of the Quaero French Medical corpus. Note that 2015 and 2016 training sets are the same (*) and that the 2016 development set is the 2015 test set (**).*

| Corpus | | Mentions | Unique mentions | Unique concepts | French coverage |
|---|---|---|---|---|---|
| EMEA 2015 | train* | 2695 | 923 | 650 | 0.67 |
| | test** | 2260 | 756 | 525 | 0.70 |
| Medline 2015 | train* | 2994 | 2296 | 1860 | 0.77 |
| | test** | 2977 | 2288 | 1847 | 0.76 |
| EMEA 2016 | train* | 2695 | 923 | 650 | 0.67 |
| | dev** | 2260 | 756 | 525 | 0.70 |
| | test | 2204 | 658 | 474 | 0.62 |
| Medline 2016 | train* | 2994 | 2296 | 1860 | 0.77 |
| | dev** | 2977 | 2288 | 1847 | 0.76 |
| | test | 3103 | 2390 | 1909 | 0.79 |

## 4. METHODS

### 4.1 Problem definition

We cast the normalization problem as a classification task. $\boldsymbol{C = \{c\}}$ is the set of all concepts $c$ (i.e., classes to predict) identified by their CUI. Each concept is associated with one semantic group $g_c$[40]. We denote the set of all concepts in a semantic group $g$ as $C_g$. A mention $\boldsymbol{m}$ is a phrase in a textual document referring to a concept. In this work, we consider these mentions to be already available and labeled with a semantic group $g_m$. The set of synonyms that share a same concept $c$ is called a synset.

For example, the concept C0678222 contains the synonyms “breast cancer”, “breast carcinoma”, “carcinoma of the breast”, is associated with the semantic type “Neoplastic Process” and is therefore in the semantic group “DISO” (Disorders). Given a french term “cancer du sein” extracted

from a document and pre-labelled with the "DISO" semantic group, our goal will be to correctly map it to the C0678222 concept.

Given a dataset, $D = [m_1, m_2, \ldots, m_n]$ our goal is to build a CUI classifier, *i.e* to learn a probability distribution $P$to predict the concept of each mention $m \in D$**:**

$$c^* = argmax_c \, P(c \mid m; \theta; H^g) \tag{1}$$

where $\theta$ represents the parameters of the encoder (detailed below), which goal is to map a mention to a dense vector space, and $H^g$ represents the embeddings of the concepts in this space, that have the same semantic group $g_m$as the mention.

## 4.2 Model

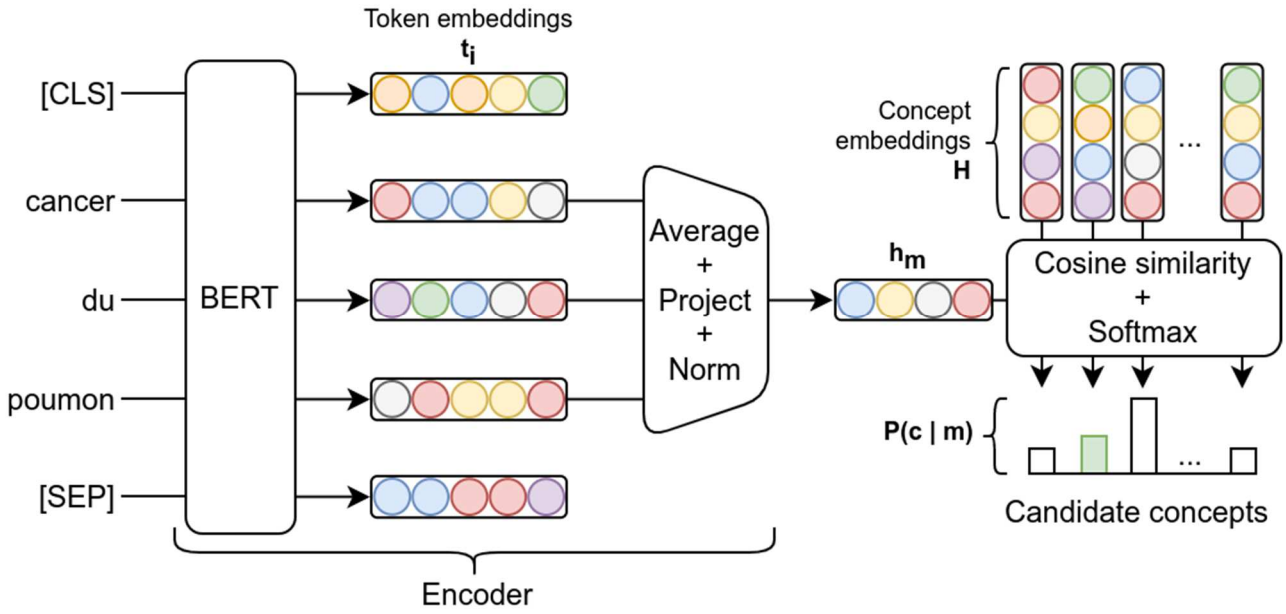


*Figure 1. Model overview. This same model is used in both steps 1 and 2 of Figure 2. During step 1, candidate concepts (bottom right of the figure) are UMLS FR + EN mirror; during step 2, candidate concepts are gold + top candidates as described in section 4.3.2.*

Our model is a classification model built on top of a sequence to sequence encoder, and we chose

pretrained Transformer [41] models in our experiments. We call this model MLNorm (for multilingual normalization). The model is described in this section and illustrated in Figure 1.

The mentions are first tokenized into wordpieces [42] and fed into a pretrained BERT encoder in order to build contextualized representations $t_i$ for each token.

$$t_i = \text{BERT(m)} \tag{2}$$

These contextualized token representations are then averaged across each mention as $\underline{t_m}$, without the first [CLS] and last [SEP] special tokens[2]

$$\underline{t_m} = \frac{1}{l-2} \sum_{i \in [1, l-1]} t_i \tag{3}$$

We then perform a projection into a lower dimension embedding to reduce the model size, apply the rectified linear activation function and normalize the result with batch normalization. This leads to a mention embedding $h_m$.

$$h_m = BN_{\mu,\sigma}\left(ReLU(W \cdot \underline{t_m} + b)\right) \tag{4}$$

where $BN_{\mu,\sigma}$is the batch normalization layer with mean $\mu$ and variance $\sigma$, and $W$ and $b$ are the projection weights and bias respectively.

Finally, we classify each mention by computing the cosine similarity between its representation and the embedding of the concepts in the semantic group of the mention. Following [33] we multiply the similarity by a hyperparameter $s$. We obtain concept probabilities by applying the softmax function on these scores.

$$P(c \mid m; \theta;\ H) = \frac{e^{s \cdot cosine\ (h_m, H_c)}}{\sum_{k \in C_g} e^{s \cdot cosine\ (h_m, H_k)}} \tag{5}$$

$$\text{where } cosine\ (h_m, H_c) = \frac{h_m}{||h_m||} \cdot \frac{H_c}{||H_c||}$$

$$H_c \text{ is the embedding of the gold concept}$$

[2] These special tokens are mandatory in a BERT model but do not carry information about the mention.

$H_k$ is the embedding of the a concept in the semantic group of $C_g$ of c

and $\theta = \{\mu, \sigma, W, b, \theta_{BERT}\}$

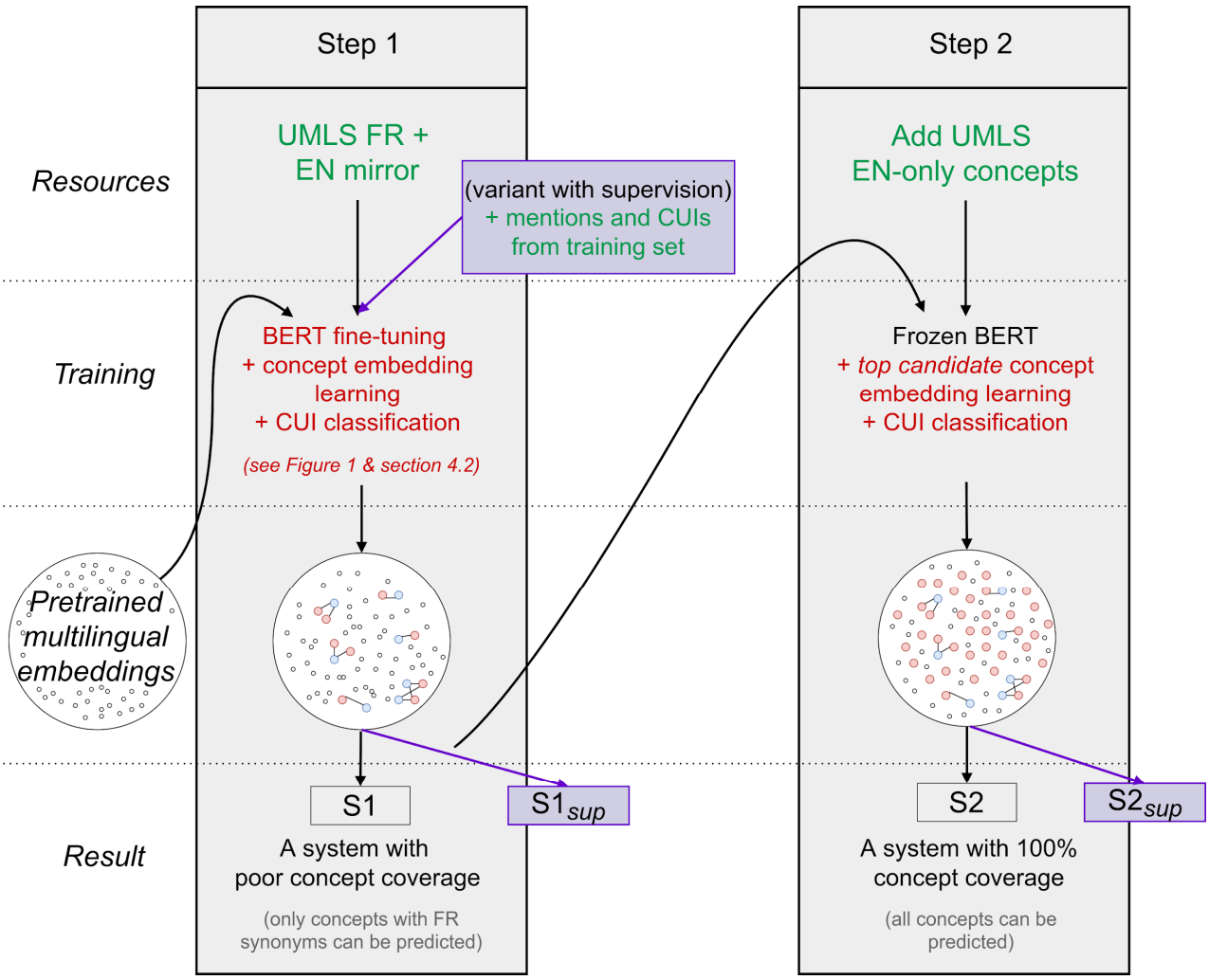


*Figure 2. Overview of our different experiments with MLNorm.*

## 4.3 Training

#### 4.3.1 Step 1: Initial full model training.

The training occurs in two steps, illustrated in Figure 2. During step 1, we learn to represent medical entities by finetuning the pretrained transformer model and train a subset of the concepts embeddings to classify synonyms from a UMLS subset.

Since this part is time-consuming and computationally expensive, we subset the UMLS and only keep the concepts with at least one French synonym (UMLS FR + EN mirror) to focus our training on the multilingual capacity of the model. This leads to a system called S1, limited to predicting only concepts having French synonyms.

#### 4.3.2 Step 2: Large-scale local concept embedding learning.

In the second phase (step 2), we freeze the transformer and projection parameters and train the representation of all the concepts (not only those having French synonyms) with a local-only learning approach. The intuition behind this is that enough synonyms were seen during step 1 so that the Transformer has an adequate representation capacity for medical mentions in general: we now just need to add the missing concepts to the model. Moreover, for each mention, most of the concepts have a near-zero probability and are not updated during the optimization (Figure 3).

For each newly-added concept, we initialize its embedding as the normalized sum of its synset's representations:

$$H_c^{mean} = \sum_{m \in synset(c)} h_m \tag{6}$$

and lookup for each synonym its $k$ highest scoring concepts as predicted by S1, that we call *top candidates*.

$$top\_candidates(m) = top_c^k \; P(c|m, \theta_1, H_1) \tag{7}$$

We train the concept embeddings as in step 1, but we consider only the batch true concepts and top candidates to compute the loss. This relates to softmax sampling methods [43] with synonym

dependent hard negatives [44]. The encoder being frozen, the synonyms' embeddings stay the same during S2, and we can efficiently compute the indices of these top candidates before starting the gradient descent of step 2. Using this method, we only have to compute gradients for a relevant subset of the concept embeddings, thus enabling a faster and more memory-efficient training.

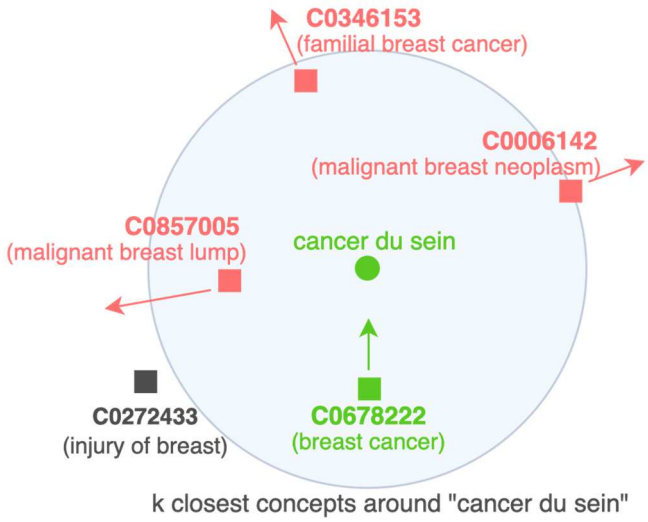


Figure 3. Overview of the local concept embedding learning. For each synonym (green dot), we precompute its k closest concepts neighbors (squares in the blue disk). Only these concept neighbors will be updated (arrows) each time the synonym appears in a batch.

We used two different learning rates, $l_{BERT}$ for the pretrained transformer, $l_{task}$ for the concept embeddings and projection layer. During the training, we vary the learning rates using two schedules. Following [45], we used a slanted triangular learning rate $l_{BERT}$ for BERT with a warm-up phase of 10% of the total number of training steps. We keep the learning rate $l_{task}$ constant during the warm-up phase and linearly decay it for the rest of the training.

### 4.4 Prediction

At inference time, a mention is tokenized and passed into the encoder and the classifier. We apply a

threshold to remove all the predictions that have a low probability.

### 4.5 Experimental setups

We performed two main sets of experiments with our model MLNorm, that we call "distantly supervised" and "fully-supervised."

In the first setup (systems S1 and S2 in Figure 2), we used only distant supervision from the UMLS, and no direct concept supervision from the available, labeled datasets from Medline and EMEA. These systems then do not suffer any potential bias from the corpus specificities and do not benefit from the redundancy of mentions within the labeled datasets. In the second setup ($S1_{sup}$ and $S2_{sup}$), we add mentions and labels from the Medline and EMEA training sets, thus enabling comparison with state-of-the-art fully-supervised approaches using these data.

Step 2 experiments use the English concepts from the five sources ("EN5") described in Section 3.1.

#### 4.5.1 Comparisons and ablation studies

**Baselines**

We compare MLNorm to the following baselines:

- the top ranked systems of respectively CLEF 2015 [18] and CLEF 2016 [4], on the same exact task of normalization from gold-standard mentions. The CLEF 2015 winning team [18] first augments the French UMLS by translating a subset of the English UMLS concepts encountered in Medline abstracts, using Google Translate. This terminology is then queried by a rule-based text indexer. The CLEF 2016 winning team [4] relies on their ECMT indexer which performs bag of words concept matching at the sentence level and integrates

up to 13 terminologies partially or totally translated into French.

- the best-performing system, to the best of our knowledge (Roller et al. [23]). They first train a local LSTM-based French to English translator on synonym pairs from the UMLS and other general domain sources. The French and English terminologies are then indexed and searched using Apache Solr through exact and fuzzy matching rules.
- as the system from Roller et al. [23], based on machine translation + English-only normalization, is quite different from our own system, we also experimented with a machine translation approach relying on our own model. This allows a fair comparison between our multilingual space learning approach and a translation-based approach. For this purpose, we translated all UMLS French terms with a state-of-the-art pretrained (*opus-mt-fr-en*) translation system [57] built with MarianMT [58] and trained on the OPUS bitext repository corpus [59]. We then trained our model with all original-English and translated-English terms. We called this strong baseline BERT-MT (using the English BERT) and mBERT-MT (using the multilingual BERT).

We also performed a range of ablation studies and additional experiments on the distantly supervised setup, in order to estimate the impact of our different choices.

**Impact of the data language**

- FR/EN: the same system without step 2 (i.e. S1 only), with the French synonyms and their English mirror
- FR-only: S1 with only the French terms
- EN-only: S1 with only the English mirror terms

**Impact of the pretrained embeddings** (our system vs. camemBERT and BERT): we compare our system with the same system using BERT embeddings trained on French data only (CamemBERT

[46], model camembert-base-uncased) or using the English-only BERT (model bert-base-uncased), instead of the multilingual BERT, in order to evaluate the contribution of the multilingual embeddings.

**Impact of more French terms** (our system vs. UMLS2019): we present an experiment using the 2019AB version of the UMLS, containing 154k concepts with French synonyms instead of 89k in the 2014AB version. With this system (UMLS2019), we aim at showing the impact of adding new concepts to the terminology used for distant supervision.

**Impact of the two-step architecture** (our system vs. 1-step): finally, we trained S1 with all the synonyms (French and English from EN5), and did not perform any step 2 with frozen embeddings. This is a much more time- and memory-consuming experiment that will allow us to estimate the quality of the model pretraining and the possible trade-off between cost and quality.

#### 4.5.2 Hyperparameter selection

For all these experiments, we trained on the training sets and evaluated on the test sets from Quaero FrenchMed datasets (Medline 2015 and 2016, EMEA 2015 and 2016). We chose the hyperparameters by selecting the best-performing values on the training set of Quaero in the distant supervision setting. We run our models on a 20 Go Tesla P40 GPU, except the 1-step experiment which required a 30Go Tesla V100 GPU.

## 5. RESULTS

As a result from the hyperparameter search described above, the token embeddings space of size 768 is projected into a space of size 350, the cosine similarity scaling parameter $s$ is 20, both dropout rates for the transformer and the projection layer are set to 0.2. We set the batch size to 128,

the maximum synonym wordpiece count to 100, and the maximum learning rates to $l_{BERT} = 2e^{-5}$and $l_{concept} = l_{proj} = 8e^{-3}$. We used Adam with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. During step 2, we preselect the $k = 100$ highest scoring concepts for each synonym. We perform the step 1 training for 15 epochs and the step 2 for 5 epochs for all models except $S1_{FR}$ and $S1_{EN}$ since they contain fewer synonyms for the same number of concepts and had to be trained longer for 30 and 20 epochs respectively. All S1 system predictions were filtered with a same threshold of 0.5 and S2 with a threshold of 0.1.

We finetuned the multilingual pretrained BERT model (*bert-base-multilingual-uncased*) in all our experiments, unless specified otherwise.

We report our main results on the test datasets from the challenges Quaero FrenchMed 2015 and 2016 (Table 3), as well as the results of our additional experiments described in previous Section (Table 4), using the traditional metrics precision, recall and F1-measure (harmonic mean of precision and recall). We also give some examples of the distantly supervised system's output (MLNorm S2), that highlight its multilingual capacities (Table 6).

*Table 3. Main results for our system on 2015 and 2016 corpora, and comparison with existing systems.*

| | | MEDLINE 2015 | | | EMEA 2015 | | |
|---|---|---|---|---|---|---|---|
| | | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| MLNorm (our system) | distantly supervised (**S2**) | 0.756 | 0.719 | 0.737 | 0.797 | 0.736 | 0.765 |
| | supervised (**S2$_{sup}$**) | 0.806 | **0.775** | **0.790** | 0.875 | **0.827** | 0.851 |
| Other supervised systems 2015 | Best system CLEF 2015 [18] | 0.805 | 0.575 | 0.671 | 1.000 | 0.774 | **0.872** |
| | Machine translation + English normalization [23] | **0.831** | 0.661 | 0.736 | **0.909** | 0.772 | 0.835 |
| | | MEDLINE 2016 | | | EMEA 2016 | | |
| | | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| MLNorm (our system) | distantly supervised (**S2**) | 0.775 | 0.734 | 0.754 | 0.746 | **0.709** | 0.727 |
| | supervised (**S2$_{sup}$**) | **0.860** | **0.740** | **0.795** | **0.832** | 0.670 | **0.743** |
| Other supervised systems 2016 | Best system CLEF 2016 [4] | 0.594 | 0.515 | 0.552 | 0.604 | 0.463 | 0.524 |
| | Machine translation + English normalization [23] | 0.771 | 0.663 | 0.713 | 0.781 | 0.692 | 0.734 |

*Table 4. Comparison of our system with a comparable machine translation approach, using our classifier.*

| Model | MEDLINE 2015 | | | EMEA 2015 | | |
|---|---|---|---|---|---|---|
| | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| **MLNorm** | **0.756** | **0.719** | **0.737** | **0.797** | 0.736 | **0.765** |
| **mBERT-MT** (MLNorm with MT) | 0.735 | 0.702 | 0.718 | 0.784 | **0.746** | **0.765** |
| **BERT-MT** (MLNorm with MT) | 0.751 | 0.698 | 0.724 | 0.774 | 0.737 | 0.755 |

*Table 5. Additional results on the 2015 corpora, about the impact of the data language, of the pretrained embeddings, the size of the French terminology and the 2-step approach.*

| Experiment | Model | MEDLINE 2015 | | | EMEA 2015 | | |
|---|---|---|---|---|---|---|---|
| | | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| Impact of the data language (step S1 only, UMLS 2014) | **FR-only** | 0.738 | 0.528 | 0.615 | 0.824 | 0.528 | 0.644 |
| | **EN-only** (EN mirror only, no FR syn) | **0.797** | 0.451 | 0.575 | **0.843** | 0.410 | 0.551 |
| | **FR/EN** (FR + EN mirror) | 0.783 | **0.621** | **0.693** | 0.827 | **0.574** | **0.678** |
| Impact of the pretrained embeddings (steps S1+ S2) | **MLNorm** | 0.756 | **0.719** | **0.737** | 0.797 | **0.736** | 0.765 |
| | **camemBERT** | **0.769** | 0.704 | 0.735 | **0.821** | 0.699 | 0.755 |
| | **BERT** (English base BERT)<br>…instead of multilingual BERT | 0.759 | 0.716 | **0.737** | 0.805 | 0.734 | **0.768** |
| Impact of more FR terms (steps S1+S2) | **MLNorm** | **0.756** | **0.719** | **0.737** | **0.797** | **0.736** | **0.765** |
| | **UMLS 2019** (instead of UMLS 2014) | 0.753 | 0.710 | 0.731 | 0.795 | 0.728 | 0.760 |
| Impact of two step training | **MLNorm** | 0.756 | **0.719** | **0.737** | 0.797 | **0.736** | **0.765** |
| | **1-step** (S1 and S2 merged in 1 step, with S2 training data) | **0.785** | 0.692 | 0.736 | **0.816** | 0.714 | 0.762 |

*Table 6. Some predictions from our system. The last two columns contain the synonyms seen during training for the target concept and the predicted one, if different. Some long or similar synonyms have been removed to improve readability.*

| System | Example mention | Expected concept and its synonyms from the train set | Predicted concept if wrong, and its synonyms for the train set |
|---|---|---|---|
| MLNorm (S2) | greffon renal | **C1261317**<br>[EN] transplanted kidney<br>[EN] kidney transplant<br>[EN] structure of transplanted kidney | ✓ |
| | cinquième métacarpien | **C0730166**<br>[EN] bone structure of fifth metacarpal<br>[EN] fifth metacarpal bone<br>[EN] fifth metacarpal | ✓ |
| | vaccination par le b.c.g | **C0199804**<br>[FR] immunisation contre la tuberculose<br>[EN] vaccination against tuberculosis<br>[EN] bcg vaccination<br>[EN] tuberculosis vaccination<br>[EN] tuberculosis immunization<br>[EN] administration of bcg vaccine<br>*… (other similar English synonyms)* | ✓ |
| | in vitro | **C0681828**<br>[EN] in vitro study<br>[EN] studies vitro<br>[EN] study vitro | **C3850137**<br>[EN] in vitro techniques<br>[EN] technique in vitro<br>[EN] in vitro as topic |
| | coffea robusta | **C0678439**<br>[EN] coffea rubusta (food) | **C1138610**<br>[EN] coffea arabica, unspecified |
| mBERT-MT | cellar<br>(translated from the French term “cave”) | **C0042460**<br>[EN] vena cava structure<br>[EN] venae cavae<br>[EN] vena cava<br>[EN] cavae<br>[EN] venae<br>[EN] vena caval structure<br>[EN] vena caval<br>[EN] vein<br>[MT] veins cellars (from “veines caves”)<br>[MT] vein cellar (from “veine cave”) | **C0007634**<br>[EN] cell<br>[EN] cell structure<br>[EN] cell type<br>[EN] cells set<br>[EN] cellula<br>[EN] cellular<br>[EN] normal cell<br>[EN] set of cells<br>[MT] cells (from “cellules”) |
| | be careful<br>(translated from the French term “attention”) | **C0004268**<br>[EN] attention<br>[EN] attentions | **C3257858**<br>[EN] my thinking is usually careful and purposeful |

## 6. DISCUSSION

Our distantly supervised system MLNorm S2 obtains very good results without concept-labeled training data (Table 3). It even reaches the same results as the best fully-supervised system published so far [23] on the corpus MEDLINE 2015 (F1=0.737 vs. 0.736) and outperforms all participants of the 2016 edition. Note that CLEF campaigns provide scores on both end-to-end task (named entity + normalization) and normalization-only task; similarly to Roller et al. [23], we compare to the latter. The much higher term redundancy can explain the better score of supervised systems on EMEA corpus (e.g., F1=0.835 and 0.734 on 2015 and 2016 for [23] vs. resp. 0.765 and 0.727 for our system S2) between training and test set (see Table 2), which gives a free boost to supervised systems but is probably not very representative of reality.

We can also see that the system using only French synonyms (FR-only) performs much poorlier, with almost 20 points less in recall than S2, which we can attribute to the missing concepts in the French UMLS.

Our experiments with translated French terms (Table 4) show that even a good machine translation model can lower the accuracy of the final model. We experimented with both English BERT and multilingual BERT to account for the impact of the transformer pre-training language. We could argue that the off-the-shelf translation model could be improved by fine-tuning on UMLS synonyms like [23]. However, we think that those results hint at the fact that translation and indexer pipeline may suffer from error cascade: being trained in an end-to-end fashion, our system does not suffer from this behavior. Table 6 shows that the ambiguity of some terms (“cave” can mean both “cellar” and “cava” in English) is lost during translation.

### 6.1 Impact of the supervision

Table 3 shows that our supervised system $S2_{sup}$ obtains a F1 gain of 5 and 9 points on resp. MEDLINE and EMEA corpora, with an improvement of both precision and recall. It outperforms other systems by a large margin on MEDLINE. It also outperforms [23] on EMEA 2015 and 2016, but not [18] that obtained a perfect precision on EMEA 2015, at the cost of many handcrafted rules and extra labeled data.

### 6.2 Impact of the data language

In Table 5, we compare the same model trained with either only the French synonyms of the UMLS 2014AB (FR-only), only the English mirror (synonyms of the same concepts, EN-only), or with both of these two sets of synonyms (FR/EN). FR/EN achieves a 8 points improvement over FR-only, despite having the same concepts coverage and the same pretrained embeddings. This indicates that a larger training set, even in a different language can help improve the system's performance by a significant margin. This improvement could be attributed to the lexical similarities between French and English languages. For example in Table 6, the only training French synonym of "*vaccination par le b.c.g*" is "*immunisation contre la tuberculose*" and shares no common word. The system can therefore benefit from the addition of similar terms, such as "*bcg vaccination*" even though they are in English.

### 6.3 Impact of the pretrained embeddings

Our experiments with French-only embeddings CamemBERT and English-only embeddings BERT, reported in Table 5, show that our hypothesis that multilingual embeddings improve the system's performance is not verified, with almost no difference between these three embeddings. French wordpieces and embeddings can handle medical terms in English, and vice versa. Even if this can be again explained in part by the proximity of the two languages concerned, the low results of EN-only, yet benefiting of much larger training data, suggest that it is not that obvious; besides, other

papers in the literature suggest that multilingual embeddings are helpful even for such pairs of languages [47, 48]. This observation may also be due to the fact that medical synonym normalization data (short word sequences) is quite different from BERT pretraining data (full sentences), so it is harder for the model to re-use its multilingual knowledge. This aspect deserves more experiments, notably on other, non-European languages. Note that biomedical-specific embeddings such as Clinical BERT [49] are not yet available in French, which is why we did not consider them. Moreover, as illustrated in Table 6, we can see that the model correctly predicts concepts, even when no common wordpieces exist between the test term and the training synonyms of the target concept. Therefore, the proximity between French and English cannot be the only explanation to the model performance. To correctly classify the mention "*cinquième métacarpien*" (fifth metacarpal bone) to its concept, without having the numeral "*cinquième*" in any of the training synonyms, the model must have learned to generalize from other concepts that contained both French "*cinquième*" and English "*fifth*" in their training synonyms.

We can also note that despite addressing out-of-vocabulary errors with wordpiece vocabularies, such errors still exist. For example in Table 6, "*robusta*" (single wordpiece "##robusta") and "*rubusta*" (two wordpieces, "##rubus" and "ta") are tokenized differently despite having almost identical characters.

### 6.4 Impact of the two-step training

Our experiment with one-step training procedure showed no improvement over the two-step training (Table 5, "1-step"), and took approximately 15 hours instead of 7 hours (5 hours for S1 and 2 hours for S2 with one million synonyms). Our two-step method can therefore effectively reduce training time without reducing accuracy by choosing an appropriate partition of the training data. Our results even show a slight loss in performance for the one-step model. This could be explained

by the regularization that occurs in the two-step training when we freeze the encoder during S2. Indeed, since most of the data seen during S2 is English, unfreezing the encoder may encourage the model to forget its inner translation capabilities.

### 6.5 UMLS 2019

Finally, our experiment with UMLS 2019AB (UMLS2019, Table 5) leads to results a little below our system, despite the much higher number of concepts having French synonyms. The system has more French terms to train on, but the coverage in Quaero corpora is not far better.

## 7. CONCLUSION AND PERSPECTIVES

We built a model using multilingual terminologies and embeddings to normalize medical concepts for a language having much lower resources than English. We obtained very good results, even in a non-fully-supervised setting, which guarantees robustness. Another attractive property of this approach is that there is no need to retrain the model from scratch when new classes (concepts) are created, which is a common problem of traditional classification methods for dynamic sets of classes.

Future work will extend these experiments to other languages and other tasks such as named entity recognition.

## ACKNOWLEDGMENTS

The authors thank the AP-HP health data warehouse for supporting this work.

## REFERENCES


[1] P. Raghavan, J. L. Chen, E. Fosler-Lussier and A. M. Lai, «How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?,» AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science, vol. 2014, p. 218–223, 2014.

[2] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler and C. G. Chute, «Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications,» Journal of the American Medical Informatics Association: JAMIA, vol. 17, pp. 507-513, 2010.

[3] A. R. Aronson and F.-M. Lang, «An overview of MetaMap: historical perspective and recent advances,» Journal of the American Medical Informatics Association : JAMIA, vol. 17, pp. 229-236, 2010.

[4] C. Cabot, R. Lelong, J. Grosjean, L. F. Soualmia and S. J. Darmoni, «Retrieving Clinical and Omic Data from Electronic Health Records,» Studies in health technology and informatics, vol. 221, p. 115, 2016.

[5] O. Bodenreider, «The Unified Medical Language System (UMLS): integrating biomedical terminology,» Nucleic Acids Research, vol. 32, pp. D267-D270, 2004.

[6] A. Névéol, C. Grouin, J. Leixa, S. Rosset and P. Zweigenbaum, «The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization,» in Proceedings of the Fourth Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing (BioTxtM2014), 2014.

[7] L. Deléger, T. Merabti, T. Lecrocq, M. Joubert, P. Zweigenbaum and S. Darmoni, «A twofold strategy for translating a medical terminology into French,» AMIA ... Annual Symposium proceedings. AMIA Symposium, vol. 2010, p. 152–156, 11 2010.

[8] P. Zweigenbaum, R. Baud, A. Burgun, F. Namer, É. Jarrousse, N. Grabar, P. Ruch, F. L. Duff, J.-F. Forget, M. Douyère and S. Darmoni, «UMLF: a unified medical lexicon for French,» International Journal of Medical Informatics, vol. 74, p. 119–124, 3 2005.

[9] K. Marko, R. Baud, P. Zweigenbaum, L. Borin, M. Merkel and S. Schulz, «Towards a multilingual medical lexicon.,» AMIA Annual Symposium proceedings. AMIA Symposium, p. 534–538, 2006.

[10] J. Grosjean, T. Merabti, B. Dahamna, I. Kergourlay, B. Thirion, L. F. Soualmia and S. J. Darmoni, «Health multi-terminology portal: a semantic added-value for patient safety.,» Studies in health technology and informatics, vol. 166, p. 129–138, 2011.

[11] A. Névéol, J. Grosjean, S. Darmoni and P. Zweigenbaum, «Language Resources for French in the Biomedical Domain,» in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, 2014.

[12] A. Névéol, H. Dalianis, S. Velupillai, G. Savova and P. Zweigenbaum, «Clinical Natural Language Processing in languages other than English: opportunities and challenges,» Journal of Biomedical Semantics, vol. 9, p. 12, 2018.

[13] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt and U. Leser, «Deep learning with word embeddings improves biomedical named entity recognition,» in Bioinformatics, 2017.

[14] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski and S. Ananiadou, «Distributional Semantics Resources for Biomedical Text Processing,» 2013.

[15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer and V. Stoyanov, «Unsupervised Cross-lingual Representation Learning at Scale,» in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 2020.

[16] F. Feng, Y. Yang, D. Cer, N. Arivazhagan and W. Wang, «Language-agnostic BERT Sentence Embedding,» 2020.

[17] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,» 10 2018.

[18] Z. Afzal, S. A. Akhondi, H. van Haagen, E. M. van Mulligen and J. A. Kors, «Biomedical Concept Recognition in French Text Using Automatic Translation of English Terms,» in CLEF 2015 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2015.

[19] J. D'Souza and V. Ng, «Sieve-based entity linking for the biomedical domain,» in ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, 2015.

[20] R. Leaman and Z. Lu, «TaggerOne: Joint named entity recognition and normalization with semi-Markov Models,» Bioinformatics, vol. 32, p. 2839–2846, 2016.

[21] R. Leaman, R. I. Doğan and Z. Lu, «DNorm: Disease name normalization with pairwise learning to rank,» Bioinformatics, vol. 29, p. 2909–2917, 2013.

[22] H. Li, Q. Chen, B. Tang, X. Wang, H. Xu, B. Wang and D. Huang, «CNN-based ranking for biomedical entity normalization,» vol. 18, p. 385, 2017.

[23] R. Roller, M. Kittner, D. Weissenborn and U. Leser, «Cross-lingual Candidate Search for Biomedical Concept Normalization,» 2018.

[24] S. Fakhraei, J. Mathew and J. L. Ambite, «NSEEN: Neural Semantic Embedding for Entity Normalization,» Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11907 LNAI, p. 665–680, 11 2018.

[25] M. Sung, H. Jeon, J. Lee and J. Kang, «Biomedical Entity Representations with Synonym Marginalization,» 2020.

[26] M. C. Phan, A. Sun and Y. Tay, «Robust representation learning of biomedical names,» in ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2019.

[27] W. Shen, J. Wang and J. Han, «Entity linking with a knowledge base: Issues, techniques, and solutions,» 2015.

[28] O. Sevgili, A. Shelmanov, M. Arkhipov, A. Panchenko and C. Biemann, «Neural Entity Linking: A Survey of Models based on Deep Learning,» 5 2020.

[29] D. P. Kingma and J. L. Ba, «Adam: A method for stochastic optimization,» 2015.

[30] S. Ioffe and C. Szegedy, «Batch normalization: Accelerating deep network training by reducing internal covariate shift,» in 32nd International Conference on Machine Learning, ICML 2015, 2015.

[31] J. Deng, J. Guo, N. Xue and S. Zafeiriou, «ArcFace: Additive Angular Margin Loss for Deep Face Recognition,» Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vols. %1 sur %22019-June, p. 4685–4694, 1 2018.

[32] F. Wang, X. Xiang, J. Cheng and A. L. Yuille, «NormFace: L2 hypersphere embedding for face verification,» MM 2017 - Proceedings of the 2017 ACM Multimedia Conference, p. 1041–1049, 2017.

[33] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li and W. Liu, «CosFace: Large Margin Cosine Loss for Deep Face Recognition,» 2018.

[34] T. Mikolov, K. Chen, G. Corrado and J. Dean, «Efficient estimation of word representations in vector space,» in 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, 2013.

[35] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, «Deep contextualized word representations,» 2 2018.

[36] E. Chiaramello, F. Pinciroli, A. Bonalumi, A. Caroli and G. Tognola, «Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes,» Journal of Biomedical Informatics, vol. 63, p. 22–32, 2016.

[37] J. Jiang, Y. Guan and C. Zhao, «WI-ENRE in CLEF eHealth Evaluation Lab 2015: Clinical

named entity recognition based on CRF,» in CEUR Workshop Proceedings, 2015.

[38] R. I. Doğan, R. Leaman and Z. Lu, «NCBI disease corpus: A resource for disease name recognition and concept normalization,» Journal of Biomedical Informatics, vol. 47, p. 1–10, 2 2014.

[39] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C. H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers and Z. Lu, «BioCreative V CDR task corpus: a resource for chemical disease relation extraction,» Database : the journal of biological databases and curation, vol. 2016, p. baw068, 5 2016.

[40] A. T. McCray, A. Burgun and O. Bodenreider, «Aggregating UMLS semantic types for reducing conceptual complexity.,» Studies in health technology and informatics, vol. 84, n° %1Pt 1, p. 216–220, 2001.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, «Attention is all you need,» Advances in Neural Information Processing Systems, p. 5999–6009, 6 2017.

[42] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes and J. Dean, «Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,» 9 2016.

[43] S. Jean, K. Cho, R. Memisevic and Y. Bengio, «On using very large target vocabulary for neural machine translation,» in ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, 2015.

[44] F. Schroff, D. Kalenichenko and J. Philbin, «FaceNet: A unified embedding for face

recognition and clustering,» in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015.

[45] C. Sun, X. Qiu, Y. Xu and X. Huang, «How to Fine-Tune BERT for Text Classification?,» vol. 11856 LNAI, 2019, p. 194–206.

[46] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah and B. Sagot, «CamemBERT: a Tasty French Language Model,» 2019.

[47] T. Pires, E. Schlinger and D. Garrette, «How multilingual is multilingual BERT?,» 2020.

[48] S. Wu and M. Dredze, «Are All Languages Created Equal in Multilingual BERT?,» 2020.

[49] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann and M. McDermott, «Publicly Available Clinical BERT Embeddings,» in Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, 2019.

[50] A R. Aronson, «Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.» in Proceedings of the AMIA Symposium, 2001

[51] I. Mondal, S. Purkayastha, S. Sarkar, P. Goyal, J. Pillai, A. Bhattacharyya and M. Gattu, «Medical Entity Linking using Triplet Network», in Proceedings of the 2nd Clinical Natural Language Processing Workshop (pp. 95–100). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019

[52] D. Wright, Y. Katsis, R. Mehta, & C.-N. Hsu, «NormCo: Deep Disease Normalization for Biomedical Knowledge Base Construction», in AKBC 2019

[53] J. Zongcheng, W. Qiang and X. Hua, «Bert-based ranking for biomedical entity normalization», in Proceedings of the AMIA Summits on Translational Science, 2020

[54] I. Lerner, N. Paris, X. Tannier. Terminologies augmented recurrent neural network model for clinical named entity recognition. Journal of Biomedical Informatics. 102, February 2020. doi: 10.1016/j.jbi.2019.103356

[55] (in French) Perceval Wajsbürt, Yoann Taillé, Guillaumé Lainé, Xavier Tannier. Participation

de l'équipe du LIMICS à DEFT 2020. in Défi Fouille de Texte (DEFT) 2020. Nancy, France, June 2020.

[56] Julien Tourille, Matthieu Doutreligne, Olivier Ferret, Nicolas Paris, Aurélie Névéol, Xavier Tannier. Evaluation of a Sequence Tagging Tool for Biomedical Texts. in Proceedings of the EMNLP Workshop on Health Text Mining and Information Analysis (LOUHI 2018). Brussels, Belgium, October 2018.

[57] Jorg Tiedemann, Santhosh Thottingal. OPUS-MT Building open translation services for the World. in Proceedings of the 22nd Annual Conference of the European Association for Machine Translation

[58] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, Andre F. T. Martins, Alexandra Birch. Marian: Fast Neural Machine Translation in C++. in Proceedings of ACL 2018, System Demonstrations

[59] Jorg Tiedemann. Parallel data, tools and interfaces in OPUS. in Proceedings of LREC, Istanbul, Turkey