Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 141-152 recibido 13-12-2020 revisado 24-01-2021 aceptado 04-02-2021

Computational Reproducibility of Named Entity Recognition methods in the biomedical domain

Reproducción computacional de métodos de reconocimiento de entidades nombradas en un dominio biomédico

Ana Garcia-Serrano,[1] Sebastian Hennig[2] and Andreas Nürnberger[2]
[1] ETSI Informatica - UNED
[2] Computer science Department - OVGU
agarcia@lsi.uned.es, sebastian.hennig@st.ovgu.de, andreas.nuernberger@ovgu.de

Abstract: Unsupervised Named Entity Recognition (NER) approaches do not depend on labelled data to function properly but rather on a source of knowledge, in which promising candidates can be looked up to find the corresponding concept. In the biomedical domain knowledge source like this already exists; namely the Unified Medical Language System (UMLS). In this paper, three different unsupervised NER models using UMLS, namely MetaMap, cTakes and MetaMapLite are evaluated and compared from the results published by Demner-Fushman, Rogers and Aronson (2017) and Reategui and Ratte (2018). The Unsupervised Biomedical Named Entity Recognition framework (UB-NER) is developed, with which the results of the experiments of the three models, five datasets and two NER tasks are presented.
Keywords: Named Entity Recognition (NER), Biomedical, supervised and unsupervised models, Unified Medical Language System.

Resumen: Los enfoques para reconocimiento de entidades nombradas no supervisados (NER, por sus siglas en inglés) no dependen de corpus con datos etiquetados, sino de una fuente de conocimiento donde buscar candidatos prometedores para encontrar el concepto correspondiente. En el ámbito biomédico existe la fuente denominada "Sistema Unificado de Lenguaje Médico" (UMLS, por sus siglas en inglés). En este artículo, se evalúan y comparan tres modelos diferentes de NER no supervisados que utilizan UMLS, a saber, MetaMap, cTakes y MetaMapLite, a partir de los resultados publicados por Demner-Fushman, Rogers y Aronson (2017) y Reategui y Ratte (2018). Para ello se desarrolla el entorno *Unsupervised Biomedical Named Entity Recognition* (UB-NER), con el que se presentan resultados de los experimentos en los modelos, cinco datasets y dos tareas NER.
Palabras clave: Reconocimiento de Entidades Nombradas (NER), Modelos biomédicos, supervisados y no supervisados, Sistema de Lenguaje Médico Unificado.

1 Introduction

The task of automated detection and the correct mapping of entities to a concept is called Named Entity Recognition (NER). Unsupervised approaches do not depend on labelled data but rather on a source of knowledge in which candidates can be looked up to find the corresponding concept. In the biomedical domain this knowledge source exists, the metathesaurus *Unified Medical Language System* (UMLS)[1], a metathesaurus in which the concepts have an associated *Concept Unique Identifier* (CUI). Three different unsupervised NER models using UMLS, namely MetaMap (Aronson, 2001), cTakes (Savova, 2010) and MetaMapLite are replicated and compared in this paper.

This research work follows the NISO Standard[2] recommendations subscribed to by

[1] https://www.nlm.nih.gov/research/umls/index.html
[2]https://www.niso.org/standards-committees/reproducibility-badging

ISSN 1989-7553. DOI 10.26342/2021-66-12

the ACM[3] to reproduce the three models in the developed framework called the *Unsupervised Biomedical Named Entity Recognition framework* (UB-NER), whose objective is to find the same results as the experiments published by Reategui and Ratte (2018) and Demner-Fushman, et al. (2017).

A section is included in the following with related work, as well as a section which describes the developed framework. Section four is devoted to the setting and description of the two kinds of experiments. A comparison of the results and some considerations on reproducibility are given when some of the configuration details are missing, unknown software versions, external resources which are no longer available or when other difficulties arise.

2 Related work

For the literature review on NER methods in the biomedical domain it can be discriminated between supervised, unsupervised and hybrid approaches (Table 1). Supervised models rely heavily on data as opposed to unsupervised models. Hence supervised approaches rely on the quality of the data and how well they represent the reality. The data needs to be labelled so that supervised models can use it for training, meaning that the model fits parameters to the underlying distribution of the data. However, the acquisition of data can usually be offset by an increased performance in contrast to unsupervised models.

Properties	**Sup.**	**UnS.**
Need for labeled data	yes	no
Domain independent	no	yes
Knowledge Source	no	yes
Arbitrary filtering of sem. types	no	yes
Restricted filtering of sem. types	yes	yes
Recognize entities	yes	yes
Metaconcept recognition	no	yes
Better accord. quality measures	yes	no
Explainability	some	yes

Table 1: Features of supervised versus unsupervised NER approaches in the biomedical domain.

Recent supervised approaches adapt the state-of-the-art approaches of neighbouring fields to the biomedical domain, giving rise to high quality NER models. For example, Lee et al. introduced BioBERT (Lee et al., 2020), a variation of the standard BERT (Devlin et al., 2019) model. The default model is additionally trained on PubMed abstracts and PubMed Central full-text articles, to fit the model to the biomedical vocabulary.

The resulting BioBERT model can solve different tasks such as NER, relationship extraction and question answering. The authors establish a new state-of-the-art performance in all three tasks. Furthermore, Cho et al. (2020) used an LSTM-CRF (Lample, 2016), to generate the embedding that is fed into the LSTM-CRF, each token goes through a bi-directional LSTM character embedding and a convolutional neural network character embedding. Instead of using the standard LSTM-CRF, the authors have inserted an attention layer between the LSTM output and the CRF, which enables the CRF to attend to the relevant parts of a sequence and put less weight on the features deemed irrelevant.

(Yu et al. 2020) published a *Generative Adversarial Network* (GAN) combined with an active learning approach, to utilize unlabelled data for training. This approach finds the different semantic types of mentions in the entity. Supervised approaches perform better in general by considering measures of quality such as precision, recall and the f1-score compared to unsupervised approaches. However, the supervised approaches rely heavily on the dataset for both the coverage of domains and the semantic filtering of the mentions.

An NER tool is considered as hybrid if it is a mixture of supervised and unsupervised methods. Supervised models may have some steps based on unsupervised methods (or vice versa), thus the model is considered hybrid. Some approximations are provided below, and some functionalities are named to show their hybrid approach. Gimli (Campos, Matos and Oliveira, 2013) is a combination of dictionary consultation and pre-processing steps usually used for unsupervised models. They use a linguistic processing tool called GDep (Sagae and Tsujii, 2007) to carry out tokenization, lemmatization, POS tagging, chunking and dependency parsing. The entities found in the dictionary consultation process are not the final output as in unsupervised settings, but rather serve as an additional feature for multiple CRF models. Another hybrid approach (Bhasuran et al. 2016) extended the CRF model

[3] https://www.acm.org/publications/policies/artifact-review-and-badging-current

and uses fuzzy matching to find rare concepts in a self-made dictionary. Instead of using one CRF model in a forward chain, they also employ a CRF model in a backward chain which reads the input sequence in reverse order. Finally in (Lara-Clares and Garcia-Serrano, 2019) a Few-Shot Learning approach is described for NER on a hybrid Bi-LSTM and Convolutional Neural Network model with four input layers to recognize multi-word entities improving precision by nearly 10%, with the addition of Wikidata entities in the vocabulary.

3 Developed Platform

This section explains the UB-NER[4] java-framework developed (Hennig, 2020). One main contribution is to bring together datasets and models and comparison functionalities in quality measures of NER experiments, to simplify its access and processing of further researchers. The second main contribution is the computational reproducibility of the most used unsupervised NER approaches (MetaMap, MetaMap Lite and cTAKES) and compare it was in (Demner-Fushman, Rogers, and Aronson, 2017) and (Reategui and Ratté, 2018), so we not include any detailed description of these approaches.

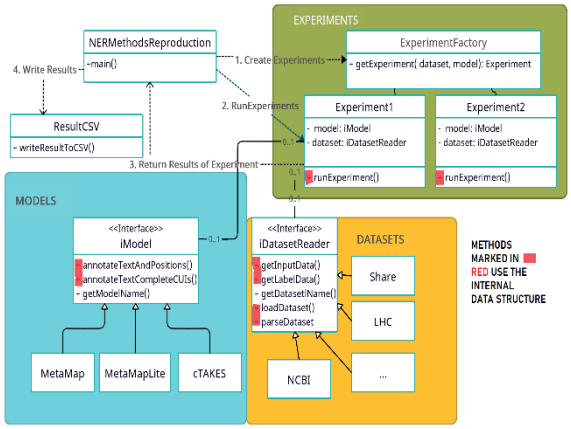

Figure 1: UB-NER High Level View. The dotted lines display the order of implementation.

All three models provide publicly available java APIs, thus facilitating the implementation of UB-NER that supports 5 datasets and 2 different NER tasks.

UB-NER consists of four components: the models, the datasets, the experiments, and the internal data structure (see Figure 1). The *annotateTextAndPositions* method first solves the NER task, giving the specific start and end position as a character offset throughout the information on the entity. The *annotateTextCompleteCUIs*, just gives a set of all entities found in the input text without any positional information. The output of the *annotateTextAndPositions* produces triplets with (start offset, end offset, concept name/ CUI). For example *'The patient has hyperlipidemia and is known to have dementia as previously stated.*' is parsed:

annotateTextAndPositions →
{(16, 30, hyperlipidemia), (45, 53, dementia)}
or {(16, 30, C0020473), (45, 53, C0497327)}
annotateTextCompleteCUIs → {C0020473, C0497327}

in which 'C0020473' is the CUI for 'hyperlipidemia' and 'C0497327' the CUI of the concept 'dementia' according to the UMLS.

Each dataset needs to implement the data set reader interface. After reading and parsing the data files, both the input and the labels can be accessed from a uniform structure. There is no pre-processing included in UB-NER because all the models implemented so far carry out the pre-processing as part of their process.

	UMLS	cTakes	Meta Map	M. M. Lite
First Experiment	2016AA	3.2.2	2016 Release	3.0
Second Experiment	2018AA[5]	3.2.0	2015 Release	-
UB-NER	2020AA	4.0.0	2020 R.	3.6.2rc5

Table 2: Versions of the UMLS and the models used in UB-NER.

A *UB-NER Experiment* is an instance of one model and one dataset, built with the *Experiment Factory*. The latest UMLS and model versions were chosen (table 2) because the two different experiments presented used different versions. Implementing a reproducible framework that automatically switches between versions would be unsuitable for the scope of this work (deviations induced by the different versions are covered in the following sections).

Apart from the semantic groups which are defined for each experiment in the configuration subsection, there are no additional configurations for MetaMap. The only

[4] Implementation technical details and reproducibility process are detailed in (Hennig and Garcia-Serrano, 2020).

[5] The UMLS version used in the original experiment is not mentioned.

additional configuration for MetaMap Lite is the segmentation method, which is set to LINES (reading each line separately instead of the complete text) for the i2b2 2010, ShARe and i2b2 2008 datasets. For the other datasets the segmentation method is not set and hence the default is used.

The *TokenProcessingPipeline* and the *FastDictionaryLookup* is used for cTAKES. Furthermore, the outputs of cTAKES are filtered to only return matches that are part of the semantic class *DiseaseDisorderMention*, since both experiments and all five datasets only contain disease and disorder references.

Although all three models contain a functionality that supports negation detection, it is not used in UB-NER, since our main goal is to reproduce the results published previously and neither of them used negation detections. However, negation can be activated by configuring the models accordingly.

The reproducible protocol is published at (Hennig, Garcia-Serrano, 2020). The framework developed is as light-weight as possible and extensible with new datasets and models following the experimental line of research established in works (Lastra and Garcia-Serrano, 2015a and b) or (Benavent et al., 2010).

4 Experiments and Evaluation

This paper's one main goal is to reproduce the two sets of NER experiments, the published by Demner-Fushman, Rogers and Aronson (2017) and the reported by Reategui and Ratte (2018). In the former, the outputs contain the name and the start and end position of each entity found. They are then compared to the gold standard and the precision, recall and f1-score are computed.

The latter experiments collate all of the entities found in a document. The entity list returned as a result is compared to a set of reference labels to locate all relevant matches. If a match is found, the candidate is added to a final output set, which is compared to the annotated gold-standard label set (subset of the reference label) and then the precision, recall and f1-score are computed.

In UB-NER each annotated concept is stored with its positional information as *AtomStringLabel*. A text usually contains more than one medical concept; hence we need a data structure to save all annotations that appear. So, each ground truth and each output consists of a set of *AtomStringLabels* and these can be compared to each other. Let L be the ground truth labels and M the labels suggested by the NER model, then

- $I = L \cap M$
- $OL = L \setminus M$
- $OM = M \setminus L$

where I is the intersection, OL are the concepts that only appear in ground truth labels and OM are the concepts that only appear in the output. These three sets can now be used to compute the *set of retrieved documents* (as $I \cup OM = M$) and the set of *relevant documents* (as $I \cup OL = L$) which are needed to calculate the precision and recall as can be seen in the following formulas. So, the calculation of OM and OL could be omitted and M and L could be used to get the retrieved and relevant document sets. In MetaMap Lite implementation OL and OM are employed for the evaluation, subsequently the precision and recall are calculated using the following formulas.

$$recall = \frac{\sum_{d \in D} |I|}{\sum_{d \in D} |\{retrieved\ documents\}|} \quad (1)$$

$$recall = \frac{\sum_{d \in D} |I|}{\sum_{d \in D} |\{relevant\ documents\}|} \quad (2)$$

$$F1 = 2\frac{precision\ .\ recall}{precision + recall} \quad (3)$$

To obtain the overall performance on a document's dataset D we do not compute the precision and recall of each document $d \in D$ and take the average, but rather accumulate all intersection set sizes and all retrieved and relevant set sizes.

In the second experiment, the multi-label classification problem, let Y be the set of all classes. Usually for a multi-label classification problem, a binary vector of size $|Y|$ for each document of D is defined, which indicates its classes. However due to the modality of the experiments, an alternate representation is used instead. For each class $y \in Y$ there is a set $L_y \subset D$, so that every $d \in L_y$ is an instance of class y. The set of leftover documents which are not an instance of y will be referred to as A_y. So, for each class $y \in Y$ there exists L and A, so that $L \subset D$, $A \subset D$ and $L \cup A = D$. We use L and A for the sets that represent the ground truth labels. Similarly, there is a set LM_y, containing all of the documents that the model predicts to be an instance of y. In the case of NER, a model

predicts a class *y* for a document *d* when an entity that is associated with *y* appears in *d*.

Let's assume MetaMap Lite is the model and we currently want to find LM for the class *Asthma*. Then we process each document $d \in D$ with MetaMap Lite. If MetaMap Lite recognizes a concept in d and assigns the CUI *C0004096* for *Asthma* to it, then d will be added to LM_{Asthma}. Thus, AM_{Asthma} is the set containing all documents in which *Asthma* is not part of the concepts detected by MetaMap Lite. At the same time, there is an LM and AM for each class $y \in Y$ where $LM \subset D$, $AM \subset D$ and $LM \cup AM = D$. Using these sets we can define the true positives (TP, entities recognized by the system that are also present in the ground truth), false negatives (FN, entities recognized that are not present in the ground truth) and false positives (FP, entities not recognized but present in the ground truth) for each class: TP = $L \cap LM$; FN = $L \setminus LM$ and FP = $LM \setminus L$. This leads to the calculation of the final score:

$$precision = \frac{|TP|}{|TP|+|FP|} \quad (4)$$

$$recall = \frac{|TP|}{|TP|+|FN|} \quad (5)$$

$$F1 = 2\frac{precision*recall}{precision+recall} \quad (6)$$

4.1 First experiments: Exact position

The four corpuses used in the exact position experiments are:

The NCBI Disease Corpus (Dogan, Leaman and Lu, 2014) consists of annotated titles and abstracts from 793 PubMed articles, annotated with MeSH and OMIM concept identifiers. As these identifiers are part of the UMLS, they can be mapped to CUIs.

Lister Hill Center (LHC) test collection is a mixture of annotated PubMed abstracts in which 150 are clinically oriented and another 150 are biology oriented. A total of 2,242 disorders are annotated and normalized to their UMLS CUIs. There exists a version of NCBI, which is also belongs to the LHC collection, that contains additional manual annotations.

The i2b2 2010 is a collection of 871 clinical notes, which provides various annotations. In this work we ignore the treatment and test annotations, following the MetaMap Lite author's evaluation strategy.

ShARe corpus contains 300 clinical notes, annotated with disorder references and normalized to a CUI if possible.

All datasets are in text-form and for each document there is a file with the text and a file with the corresponding annotated entities. The authors of the original experiments parsed the labels to brat standoff format[6] and the CUIs are omitted as the preferred concept names, the human readable identifier in UMLS, are used for comparison as they can be interchanged. They compare the concept name as well as the start and the end positions in the text. In this work the labels are not parsed to the brat standoff format, but the concept name and offsets are equally compared using the *AtomStringLabel* format.

A typical label could look like *"lung cancer | 14 | 25*", in which lung cancer is the preferred name, 14 is the number of the starting character and 25 the ending one. The character offsets are all relative to the first character of the document. Each label has a semantic type assigned to it and, following the work to be reproduced, we only consider labels that fall under one of the semantic types *Disorder* or *GeneralDisorder*. The main reasons for this choice are that the datasets and tools are heavily skewed toward these semantic types and also because of their importance in clinical text processing and downstream applications, such as the extraction of phenotypes or adverse reactions to drugs (Segura-Bedmar and Martínez, 2017).

As mentioned before, the concept names are used for the gold-standard labels instead of the CUIs. Therefore, in this work we need the output of the models to be a concept name, too (to make then iniformal). Each of the three models MetaMap, MetaMap Lite and cTakes, can output the multiple formats of a found concept. Namely the CUI, the preferred concept name as saved in the dictionary (UMLS) and the concept name found in the text. We decided to use the concept name found in the text, which is also used in the ground-truth labels. Using the preferred concept names as defined in a dictionary, would lead to problems in assigning correct offsets in the model outputs as well as in the labels, since the length of the dictionary entry can vary from the length of the corresponding phrase found in the text.

Although it is not mentioned in the original paper, but directly influences the results, the

[6] http://brat.nlplab.org/standoff.html

MetaMap and MetaMap Lite output is restricted to a list of semantic types. The nine semantic types mentioned in the code kindly provided by the authors of MetaMap Lite are: *Acquired Abnormality (acab), Congenital Abnormality (cgab), Injury or Poisoning (inpo), Pathologic Function (patf), Disease or Syndrome (dsyn), Anatomical Abnormality (anab), Neoplastic Process (neop), Mental or Behavioral Dysfunction (mobd), Sign or Symptom (sosy)*. Restricting the output to these semantic types increases the precision of MetaMap and MetaMap Lite, since any entities found that are not annotated in the gold standard as disease, e.g. entities of the semantic type plant, are discarded.

4.2 Second experiments: Classification

The authors (Reategui and Ratte, 2018) ran two experiments identifying whether a comorbidity[7] is present or not in a discharge summary. They differ in the labels used. In the first experiment (Single CUI experiment), single UMLS concepts are assigned to each comorbidity which should be predicted by the models. For the Multiple CUIs experiment, additional UMLS concepts are added to some of the comorbidities thus forming an aggregation of CUIs. This task is easier since the models only need to find one of the CUIs mentioned in a concept aggregate of a comorbidity to get a successful match.

The i2b2 2008 obesity challenge dataset used in this experiment (Uzuner, 2009) contains 1,237 medical discharge summaries of obese and diabetic people. It is annotated with 15 possible comorbidities of obesity. The labels indicated for each comorbidity in the underlying medical record are: **present** (the patient has/had the disease); **absent** (the patient does/did not have the disease); **questionable** (the patient may have the disease) and **unmentioned** (the disease is not mentioned in the discharge summary).

Aiming exactly at reproducing the results of the authors, we selected the subset of 412 summaries which had obesity as a comorbidity and the annotated gold standard was taken and changed into a binary classification task. We discriminate between present and absent, where a comorbidity is present if and only if it is tagged as present in the gold standard. If it is tagged as either absent, questionable, or unmentioned we consider it as absent. With this new binary presentation two sets of documents can be created for each comorbidity, namely *L* and *A*, as mentioned in section 4. In the original experiments of Reategui and Ratte (2018), *D* corresponds to the set of all 412 obesity discharge summaries, and the set of classes of comorbidities (*Y)* considered are: *Pathologic Function (patf)*; *Disease or Syndrome (dsyn); Therapeutic* or *Preventive Procedure (topp); Mental or Behavioral Dysfunction (mobd)*. We refer to the original publication for explanations on the aggregation process and the reasoning behind the choices for the aggregations used in the second experiment.

The two experiments are carried out using the precision and recall calculations stated in section 4. The only difference is the creation of the *LM* sets for the Multiple CUIs experiment. In the Single CUI experiment, a summary is only part of $LM_{\text{Depression}}$, if the model detects an entity with CUI *C0011570* in it. For the Multiple CUI experiment it is enough for a summary to be included in $LM_{\text{Depression}}$, if the model manages to detect either the concept *C0011570* or *C0011581*.

Since the two classification experiments described are different from the first one based on the work of the MetaMap Lite, we have created an additional *ExperimentCompleteDoc* class in UB-NER, where instead of looking at each concept found separately, we create a list of all concepts found. The resulting list is checked against the 14 available comorbidities considered (*Hypertriglyceridemia* was excluded due to a lack of sufficient examples). If a comorbidity is found in the document, it is added to the *LM* set.

Changes were also needed in the dataset loading. Instead of creating *AtomStringLabels* for each document, we assigned a document to the *L* set if a comorbidity was annotated as present in the gold standard. The scores are then computed after all *L* and *LM* sets are calculated.

5 Results Comparison

In this section the results obtained by UB-NER reproducing the two original experiments are compared with the results published. Furthermore, the delta between the two evaluations is calculated by subtracting the original from UB-NER score. Hence a positive

[7] Comorbidity refers to the presence of more than one disorder (co-existing) in the same person.

entry means that our model is better than the original and a negative means that is not.

5.1 Exact Position Experiments Results

We found similar results (table 3) as published in the original article as set out in table 4. MetaMap Lite outperforms the others in terms of precision, recall and f1 and only on the ShARe dataset, MetaMap marginally beat MetaMap Lite.

In the original experiment the *AggregatePlaintextUMLSProcessor* were used in the cTAKES pipeline. Unfortunately, we could not run it since it took more time to process one datapoint, than it took MetaMap Lite to process the complete dataset. Hence, we used the fast pipeline provided by cTAKES.

Datasets	MetaMap		
	P	R	F1
LHC NCBI	0.546	0.583	0.564
LHC-Bio Cits	0.396	0.608	0.479
ShARe	0.444	0.662	**0.532**
LHC-Clin Cits	0.561	0.635	0.596
i2b2 2010	0.364	0.347	0.355

Datasets	MetaMapLite		
	P	R	F1
LHC NCBI	0.664	0.714	**0.688**
LHC-Bio Cits	0.468	0.711	**0.564**
ShARe	0.483	0.585	0.529
LHC-Clin Cits	0.635	0.711	**0.671**
i2b2 2010	0.395	0.349	**0.371**

Datasets	cTAKES		
	P	R	F1
LHC NCBI	0.483	0.607	0.538
LHC-Bio Cits	0.443	0.549	0.490
ShARe	0.464	0.417	0.440
LHC-Clin Cits	0.517	0.549	0.533
i2b2 2010	0.315	0.202	0.246

Table 3 (a), (b), (c) UB-NER results on the exact position experiments.

Most of the deviations detailed in table 4 (on average our scores are 0.035 worse than originals) can be explained by the differences in comparing the model output to the label set. We parsed the labels and the model output to *AtomStringLabel*, whereas in the original experiments the brat standoff format was used.

There are some cases in which the output of UB-NER identifies the positions correctly, but the entity name does not exactly match the label name. For example, "524 555 glucose/galactose malabsorption" is the output and "524 555 glucose malabsorption" is the gold standard. Parsing the model output to the brat standoff, changes those cases to be mapped correctly. However, only 0.34% of all labels are affected.

	METAMAP		
DATASETS	P	R	F1
LHC NCBI	-0.057	-0.1	-0.077
LHC-BIO CITS	-0.072	-0.148	-0.099
SHARE	-0.151	0.181	0
LHC-CLIN CITS	-0.027	-0.137	-0.072
I2B2 2010	-0.017	-0.01	-0.013

	METAMAP LITE		
DATASETS	P	R	F1
LHC NCBI	-0.067	-0.005	-0.037
LHC-BIO CITS	-0.207	-0.068	-0.16
SHARE	-0.259	0.164	-0.009
LHC-CLIN CITS	-0.059	-0.038	-0.029
I2B2 2010	-0.075	0.03	-0.009

	CTAKES		
DATASETS	P	R	F1
LHC NCBI	0.013	0.069	0
LHC-BIO CITS	-0.028	-0.057	-0.04
SHARE	0.001	-0.045	-0.022
LHC-CLIN CITS	0.091	-0.05	0.035
I2B2 2010	-0.004	-0.139	-0.083

Table 4 (a), (b), (c): Delta to Original Results.

MetaMap and MetaMap Lite differ in precision and recall on the ShARe dataset, but in such proportions that they offset each other and the f1 score stays the same. The ShARe dataset does not have the name of the entity as a label, but instead each entity is tagged with its CUI. MetaMap Lite converted those CUI labels to the brat standoff format. In UB-NER the outputs of the models were adapted, mapping the entity to the corresponding CUI, allowing the output to be matched against the gold labels given by the ShARe dataset, containing positional information and the CUI.

The differences of MetaMap for the LHC-Bio Cits and LHC-Clin Cits are induced by the aggregation of variations from the original experiment. In addition to the differences between the brat standoff and the *AtomStringLabel*, the output of MetaMap is also different. The original experiment uses the fielded MetaMap (MMI) output. Unfortunately, the MetaMap API does not support this output format. We approximate the MMI output as closely as possible with the API available tools.

However, there are limits which cannot be easily overcome. For example, the phrase "*transposition of the great vessels*" is recognized

as *Transposition of Great Vessels* and CUI *C0040761* when the fielded MMI output is used. When the API is used, two independent concepts, namely *Transposition* with CUI *C0040759* and *Great vessels* with CUI *C0225991* are returned by MetaMap from which the latter is removed from the output since the semantic type of *Great vessels* is not part of the list used for the experiment.

Therefore, without adapting the semantic types, it is not possible to get the same output with the API as the console version with a fielded MMI output. Both output forms identify abbreviations but only the fielded MMI output returns the short form mentioned in the text, which is also the one used in the labels most of the time. The API on the other hand, only returns the full name of the corresponding concept instead of the abbreviation. A heuristic is implemented in UB-NER to map those complete matches back to the abbreviations found but is unable to produce the same output as the fielded MMI.

Changes in the UMLS versions also causes some entities, that were previously found, to no longer be recognized. For example the concept *HIV* is part of the semantic group *Disease or Syndrome (dysn)* in former UMLS versions, while the current version of the UMLS used in UB-NER maps HIV to the semantic group *Virus (virs)* which is not part of that list.

Theoretically these problems are present in all datasets processed by our implementation of MetaMap. The greater influence of these factors on the LHC-Bio Cits and LHC-Clin Cits among others stems from the fact that these datasets are relatively small compared to the other, and hence single errors have a greater impact on the overall score. The deviation of the precision of MetaMap Lite on the LHC-Bio Cits is because MetaMap Lite was able to recognize many more abbreviations with the UMLS 2020AA than with older versions.

Unfortunately, the texts in the LHC-Bio Cits dataset contain a lot of abbreviations for phrases that are not diseases. For example, the phrase "*Corticotropin-releasing factor (CRF)",* where the abbreviation *CRF* is used for all other occurrences in the text, is identified by MetaMap Lite with the UMLS 2020AA. Even though the correct concept *C0772289* belonging to this phrase, can be found by MetaMap Lite, the resulting semantic type for this match is not contained in the list of accepted semantic types. This would result in the *CRF not being matched*. Unfortunately, the abbreviation CRF is also used for the concept *Cancer-related fatigue* with CUI *C4274302*. Hence MetaMap Lite outputs a wrong interpretation of CRF.

Naturally biological abstracts contained in the LHC-Bio Cits dataset, also contain abbreviations for biological phrases, and some concepts are mapped to the same abbreviation, even though they are completely uncorrelated. These false positives, who's weighting to the total score is enhanced by the fact that an abbreviation is frequently used, results in a lower precision. So, while it is a good idea to include abbreviations to increase recall, it can decrease the precision disproportionately.

5.2 Classification Experiments Results

The best model for each comorbidity shows that MetaMap Lite cannot outperform MetaMap and cTakes, in contrast to results in previous section, but it can match their performance.

Precision and recall of this task are higher than in the exact position because: (1) No position tagging is required; (2) The task is aligned with the dataset: nearly all biomedical entity mentions belong to one of the 14 target concepts; and (3) The entities are not verified one by one but count as a match if the entity appears at least once in the document.

In general, our UB-NER results (tables 5, 6 and 7) match closely with the results of the original work and differences can be attributed to the use of different versions of UMLS and that: (1) no configuration details of MetaMap are given, (2) neither are cTAKES and (3) neither was the UMLS version mentioned for MetaMap nor cTAKES.

	MetaMap					
	Single CUI Exp.			Multiple CUI Exp.		
Entity	P	R	F1	P	R	F1
CHF	0.864	0.927	**0.895**	0.864	0.921	0.891
Hypertension	0.95	0.97	0.96	0.949	0.967	0.958
Venous Insufficiency	1	0.316	0.48	1	0.316	0.48
Gout	0.945	0.963	0.954	0.945	0.963	0.954
CAD	0.839	0.672	0.746	0.821	0.688	0.749
Gallstones	0.982	0.7	0.818	0.97	0.8	0.877
Depression	0.932	0.932	**0.932**	0.931	0.92	**0.926**
Asthma	0.885	0.906	0.895	0.884	0.894	0.889
GERD	0.911	0.947	0.929	0.911	0.947	0.929
OA	0.866	0.816	0.84	0.866	0.816	**0.84**
Hypercholesterolemia	0.935	0.571	**0.709**	0.948	0.84	**0.891**
Diabetes	0.885	0.686	0.773	0.888	0.895	0.892
OSA	0.924	0.758	**0.833**	0.923	0.75	**0.828**
PVD	0.974	0.974	**0.974**	0.974	0.974	**0.974**
Average	0.921	0.796	0.838	0.92	0.835	0.863

Table 5: Results for MetaMap Experiments.

MetaMap Lite

	Single CUI Exp.			Multiple CUI Exp.		
Entity	P	R	F1	P	R	F1
CHF	0.873	0.915	0.893	0.873	0.915	**0.893**
Hypertension	0.95	0.976	0.963	0.95	0.976	0.963
Venous Insufficiency	1	0.316	0.48	1	0.316	0.48
Gout	0.944	0.944	0.944	0.944	0.944	0.944
CAD	0.826	0.892	0.858	0.81	0.892	0.849
Gallstones	0.979	0.588	0.734	0.966	0.7	0.812
Depression	0.761	0.943	0.843	0.761	0.943	0.843
Asthma	0.892	0.871	0.881	0.892	0.871	0.881
GERD	0.913	0.961	**0.936**	0.913	0.961	**0.936**
OA	0.897	0.598	0.717	0.897	0.598	0.717
Hypercholes-terolemia	0.949	0.537	0.686	0.959	0.811	0.879
Diabetes	0.886	0.725	0.797	0.885	0.899	0.892
OSA	0.919	0.711	0.802	0.919	0.711	0.802
PVD	0.97	0.842	0.901	0.97	0.842	0.901
Average	0.911	0.772	0.817	0.91	0.813	0.842

Table 6: MetaMap Lite Experiments Results.

cTAKES

	Single CUI Exp.			Multiple CUI Exp.		
Entity	P	R	F1	P	R	F1
CHF	0.924	0.661	0.77	0.924	0.661	0.77
Hypertension	0.943	0.991	**0.966**	0.943	0.991	**0.966**
Venous Insufficiency	1	0.316	0.48	0.633	1	**0.776**
Gout	0.945	0.963	0.954	0.945	0.963	0.954
CAD	0.828	0.903	**0.864**	0.828	0.903	**0.864**
Gallstones	0.984	0.788	0.875	0.959	0.888	**0.922**
Depression	0	0	0	0.719	0.989	0.833
Asthma	0.867	1	**0.929**	0.867	1	**0.929**
GERD	0.862	0.987	0.92	0.862	0.987	0.92
OA	0.906	0.667	0.768	0.906	0.667	0.768
Hypercholes-terolemia	0.941	0.549	0.693	0.954	0.829	0.887
Diabetes	0.888	0.826	**0.855**	0.877	0.915	**0.896**
OSA	0.921	0.727	0.812	0.921	0.727	0.812
PVD	0.969	0.816	0.886	0.969	0.816	0.886
Average	0.856	0.728	0.769	0.879	0.881	0.87

Table 7: UB-NER results for cTAKES experiments.

The greatest discrepancy in table 8 is the single CUI *Depression* which is mapped to the CUI *C0011570*. cTAKES maps all occurrences of *C0011581*. In the multiple CUIs experiment in which the CUI *C0011581* is added, the results match up again for cTAKES and a marginal improvement in precision is achieved.

The MetaMap implementation used in UB-NER also gave rise to better precision results in the single CUI and multiple CUIs experiment. In the literature, *Depression* is hard to recognize correctly, because usually refers to a mental disorder, but in the biomedical domain it can also refer to a "reduction".

There is also a significant difference for *Atherosclerotic Cardiovascular Disease (CAD)* in the single CUI experiment, whereas the difference in the multiple CUIs experiment is negligible. In former UMLS versions, instances of CAD were solely mapped to the concept *Coronary arteriosclerosis*, CUI *C0010054*, however in the current version it has a new one, the *Coronary artery disease* with CUI *C1956346*. The CUI mapping table shows that *C1956346* was used in the Single and *C0010054* was added for the Multiple experiment.

	MetaMap			cTAKES		
Entity	P	R	F1	P	R	F1
CHF	-0.006	0.037	0.015	0.064	**-0.259**	**-0.12**
Hypertension	-0.01	-0.02	-0.02	-0.007	0.001	-0.004
Venous Insufficiency	0	0.026	0.04	0	0.026	0.04
Gout	-0.005	-0.017	-0.006	-0.005	-0.017	-0.006
CAD	-0.021	**0.222**	**0.156**	-0.012	-0.017	-0.016
Gallstones	-0.018	-0.03	-0.022	-0.016	0.008	-0.005
Depression	**0.232**	0.042	**0.142**	**-0.71**	**-0.99**	**-0.82**
Asthma	-0.015	-0.024	-0.015	-0.013	0	-0.011
GERD	0.021	-0.023	-0.001	-0.018	-0.003	-0.01
OA	-0.004	0.056	0.03	-0.044	-0.003	-0.012
Hypercholes-terolemia	-0.005	-0.019	-0.021	-0.009	0.039	0.033
Diabetes	-0.025	0.036	0.013	-0.022	-0.004	-0.015
OSA	-0.016	-0.022	-0.017	-0.019	-0.033	-0.028
PVD	0.004	0.004	0.004	-0.001	-0.024	-0.014

Table 8: Delta Single CUI.

	MetaMap			cTAKES		
Entity	P	R	F1	P	R	F1
CHF	-0.006	0.031	0.011	0.064	**-0.259**	**-0.12**
Hypertension	-0.011	-0.023	-0.022	-0.007	0.001	-0.004
Venous Insufficiency	**0.296**	**-0.589**	**-0.312**	-0.067	0	-0.048
Gout	-0.005	-0.017	-0.006	-0.005	-0.017	-0.006
CAD	-0.009	0.088	0.059	0.018	-0.017	-0.006
Gallstones	-0.02	-0.065	-0.043	-0.011	-0.002	-0.008
Depression	**0.225**	-0.01	**0.124**	0.009	-0.001	0.013
Asthma	-0.016	-0.036	-0.021	-0.013	0	-0.011
GERD	0.021	-0.023	-0.001	-0.018	-0.003	-0.01
OA	-0.004	0.056	0.03	-0.044	-0.003	-0.012
Hypercholes-terolemia	-0.012	-0.04	-0.029	-0.006	0.019	0.007
Diabetes	-0.022	0.005	-0.008	-0.013	-0.005	-0.014
OSA	-0.017	-0.03	-0.022	-0.019	-0.033	-0.028
PVD	0.004	0.004	0.004	-0.001	-0.024	-0.014

Table 9: Delta Multiple CUIs.

Table 9 shows that *Venous Insufficiency* has significant differences for the multiple CUIs experiment. This stems from the addition of the concept *Postthrombotic syndrome* with CUI *C0277919*. In the former versions of the UMLS, *venous stasis* is mapped to the CUI *C0277919*, which explains the improved performance in the original. In the 2020AA version of the UMLS a new concept for *venous stasis* was introduced with CUI *C441518*. Hence all instances that were previously mapped to *C0277919* are now mapped to *C441518*. If we

substituted the *C0277919* with *C441518*, we would likely get the same results.

The discrepancy in the single CUI and multiple CUIs experiment for *CHF* in cTAKES is also brought about by software versions. In the current one, like *Venous insufficiency*, instances that can be mapped to more specific concepts are no longer mapped to the general concept *C0018802*, resulting in a lower recall. It is necessary to have notice that even if the three models were processed by the UB-NER for this experiment and results explained, MetaMap Lite it is not shown in delta tables 7 and 8 because it was not included in the comparison of the original work in (Reategui and Ratté, 2018, thus it is not possible to calculate any delta for MetaMap Lite.

5.3 Computational reproducibility

UB-NER was able to reproduce the results published in (Demner-Fushman, Rogers, and Aronson, 2017) and (Reategui and Ratte, 2018) with no significant differences according to the student's t-test, proving the published findings as reproducible and correct.

For the student's t-test, the p-value is computed by using a two-sided t-distribution on two paired sample sets. Our null hypothesis H_0 states that the average performance of the compared implementations is equal, whilst the alternative hypothesis states that their average performance is different. We choose a 5% significance level and say that the performance differs significantly if we must reject H_0 i.e. the p-value is smaller than 0.05. On the other hand, if the p-value is larger or equal to 0.05 H_0 holds and the differences in performance are considered insignificant.

If we take the values from table 3 and the original results from in (Demner-Fushman, Rogers, and Aronson, 2017) the calculation of the p-value yields 0.198 and hence shows that our results are not significantly different from the original results. Analogous the p-value for the Single CUI experiment (table 5) is 0.199 and 0.373 for the Multiple CUI experiment (table 6) respectively, indicating that the differences to the results published in (Reategui and Ratte, 2018) are also not significant.

6 Conclusions

The two NER in the biomedical domain widely used are the following unsupervised models: MetaMap with just a few supervised parts in its pipeline (the POS-tagger) and cTAKES which has more pre-trained supervised parts in its pipeline. Both use the UMLS to identify and extract medical entities from text and were compared in (Reategui and Ratté, 2018) showing very similar behaviour using the i2b2 2008 dataset. In (Demner-Fushman, Rogers, and Aronson, 2017) MetaMap and cTAKES were compared with MetaMap Lite, a Java implementation of MetaMap focusing on real-time processing speed.

We presented the UB-NER framework to validate published results in the original comparisons of (Demner-Fushman, Rogers, and Aronson, 2017) and (Reategui and Ratté, 2018), with a discussion justifying the differences found and explaining how the different versions of UMLS, the abbreviations considered and other related features, impact on the results.

UB-NER enables the computational reproduction of scientific research results, bringing together biomedical datasets and models for NER models, so removing barriers in the dataset access and NER processing to the researchers, i.e. all models in the original papers have different input/output formats and not in UB-NER. To configure an experiment in UB-NER you only must do some database and model selection to obtain results and quality measures.

We plan to extend UB-NER to support more datasets, models, and experiments for unsupervised as well as supervised approaches. Furthermore, we want to create a novel NER method that uses a supervised approach, exploiting additional information provided by UMLS, to enhance the usability of entities found for downstream tasks.

Acknowledgements

Thanks to Juan J. Lastra-Díaz y Alicia Lara-Clarés for their initial comments. We also want to thank Dina Demner-Fushman and Willie Rogers. The feedback provided by them were really helpful.

Bibliography

Aronson, A.R. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: theMetaMap program. *Proc. AMIA Annual Symposium*, pages 17–21, ISSN 1531605X.

Benavent, J., X. Benavent, E. de Ves, R. Granados, and A. Garcia-Serrano. 2010. Experiences at ImageCLEF 2010 using CBIR and TBIR Mixing Information Approaches. CLEF CEUR-WS, vol 1176.

Bhasuran, B., G. Murugesan, S. Abdulkadhar, and J. Natarajan. 2016. Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. *Journal of Biomedical Informatics* 64 (Dec), pp. 1–9. doi: 10.1016/j.jbi. 2016.09.009.

Campos, D., S. Matos, and J. L. Oliveira. 2015. Gimli: Open source and high-performance biomedical name recognition. *BMC Bioinformatics* 14.1 Feb, p. 54. doi: 10.1186/1471-2105-14-54

Cho, M., J. Ha, C. Park, and S. Park. 2020. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. *Journal of Biomedical Informatics* 103 (Mar) p. 103381. doi: 10.1016/j. jbi.2020.103381.

Demner-Fushman, D., W. J. Rogers, and A. R. Aronson. 2017. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *J. of the American Medical Informatics Association* 24.4, pp. 841–844. doi: 10.1093/jamia/ocw177.

Devlin, J., M.Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Technical report. https://github.com/tensorflow/tensor2tensor.

Dogan, R.I., R. Leaman, and Z. Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, doi: 10.1016/j.jbi.2013.12.006.

Hennig, S. 2020. An experimental survey of Named Entity Recognition methods in the biomedical domain. Master Data and Knowledge Engineering. Faculty of Computer Science. OVGU. A. Garcia-Serrano and A. Nürnberger supervisors.

Hennig, S. and A. Garcia-Serrano. 2020. Reproducible experiments on the master thesis: An experimental survey of Named Entity Recognition methods in the biomedical domain, UNED *e-cienciaDatos, V1* (dec) https://doi.org/10.21950/DYAZRE.

Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. 2016. Neural architectures for named entity recognition. *Proc. of NAACL HLT 2016*, pp. 260–270.

Lara-Clares, A., A. Garcia-Serrano. 2019. LSI2_UNED at eHealth-KD Challenge 2019: A Few-shot Learning Model for Knowledge Discovery from eHealth Documents. *CEUR-WS*, vol 2421, IberLEF. Bilbao, Spain.

Lastra-Díaz, J.J. and A. Garcia-Serrano. 2015a. A novel family of IC-based similarity measures with a detailed experimental survey on WordNet. *Engineering Applications of Artificial Intelligence* 46, 140-153.

Lastra-Díaz, J.J. and A. Garcia-Serrano. 2015b. A new family of information content models with an experimental survey on WordNet. *Knowledge-Based Systems* 89, 509-526.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C.Ho So, and J. Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36.4 (Feb), pp. 1234–1240. doi: 10.1093/bioinformatics/ btz682.

Merkel, D. 2014. Docker: lightweight Linux containers for consistent development and deployment. https://dl.acm.org/doi/10.5555/ 2600239.2600241.

Mowery, D. 2013. ShAReCLEF eHealth Evaluation Lab 2014 (Task 2): Disorder Attributes in Clinical Reports. *PhysioNet* https://doi.org/10.13026/0zgk-9j94.

Reategui, R. and S. Ratte. 2018. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Medical Informatics and Decision Making* 18.3, p. 74. doi: 10.1186/s12911-018-0654-2.

Sagae, K. and J. Tsujii. 2007. Dependency Parsing and Domain Adaptation with LR

Models and Parser Ensembles. In *Proc. of the EMNLP-CoNLL, 2007*, pp. 1044–1050 https://www.aclweb.org/anthology/D071111

Savova, G., J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical* Informatics Association 17(5):507–513.
DOI: 10.1136/jamia.2009.001560

Segura-Bedmar, I. and P. Martínez. 2017. Simplifying drug package leaflets written in Spanish by using word embedding. *Journal of Biomedical Semantics* 8, 45. https://doi.org/10.1186/s13326-017-0156-7

Uzuner, A. 2009. Recognizing Obesity and Comorbidities in Sparse Data. *Journal of the American Medical Informatics Association*, 16(4):561–570, 7.

Gang, Y., Y. Yang, X. Wang, H. Zhen, G. He, Z. Li, Y. Zhao, Q. Shu, and L. Shu. 2020. Adversarial active learning for the identification of medical concepts and annotation inconsistency. *Journal of Biomedical Informatics 108* (Aug), p. 103481.
https://doi.org/10.1016/j.jbi.2020.103481.