ICD-10 coding of death certificates with the NCBO and SIFR Annotators at CLEF eHealth 2017

Andon Tchechmedjiev, Amine Abdaoui, Vincent Emonet, Clement Jonquet

ICD-10 coding of death certificates with the NCBO and SIFR Annotators at CLEF eHealth 2017

Andon Tchechmedjiev[1], Amine Abdaoui[1], Vincent Emonet[1], and Clement Jonquet[1,2]

(1) Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM)
University of Montpellier & CNRS, France
(2) Center for BioMedical Informatics Research (BMIR)
Stanford University, USA
{prenom.nom}@lirmm.fr

Abstract. The SIFR BioPortal is an open platform to host French biomedical ontologies and terminologies based on the technology developed by the US National Center for Biomedical Ontology (NCBO). The portal facilitates the use and fostering of terminologies and ontologies by offering a set of services including semantic annotation. The SIFR Annotator (http://bioportal.lirmm.fr/annotator) is a publicly accessible, easily usable ontology-based annotation tool to process French text data and facilitate semantic indexing. The web service relies on the ontology content (preferred labels and synonyms) as well as on the semantic of the ontologies (is-a hierarchies) and their mappings. The SIFR BioPortal also offers the possibility of querying the original NCBO Annotator for English text via a dedicated proxy that extends the original functionality. In this paper, we present a preliminary performance evaluation of the generic annotation web service (i.e., not specifically customized) for coding death certificates i.e., annotating with ICD-10 codes. This evaluation is done against the CépiDC/CDC CLEF eHealth 2017 task 1 manually annotated corpus. For this purpose, we have built custom SKOS vocabularies from the CéPIDC/CDC dictionaries as well as training and development corpora, for all three tasks using a most frequent code heuristic to assign ambiguous labels. We then submitted the vocabularies to the NCBO and SIFR BioPortal and used the annotation services on task 1 datasets. We obtained, for our best runs on each corpus the following results: English raw corpus (69.08% P, 51.37% R, 58,92% F1); French raw corpus (54.11% P, 48.00% R, 50,87% F1); French aligned corpus (50.63% P, 52.97% R, 51.77% F1).

Keywords: Semantic annotation, SIFR Annotator, NCBO Annotator, ICD-10 coding, Biomedical ontologies.

1 Introduction

Biomedical data integration and semantic interoperability is necessary to enable translational research. The biomedical community has turned to ontologies and

terminologies to describe their data and turn them into structured and formalized knowledge [1,2]. Ontologies help to address the data integration problem by playing the role of common denominator. One way of using ontologies is by means of creating semantic annotations. An annotation is a link from an ontology term to a data element, indicating that the data element (e.g., article, experiment, clinical trial, medical record) refers to the term [3]. When doing ontology-based indexing, we use these annotations to "bring together" the data elements from these resources.

The community has turned toward ontologies to design semantic indexes of data that leverage the medical knowledge for better information mining and retrieval. Despite a large adoption of English in science, a significant quantity of biomedical data uses the French language. Besides the existence of various English tools, there are considerably less terminologies and ontologies available in French [4] and there is a strong lack of related tools and services to exploit them. This lack does not match the huge amount of biomedical data produced in French, especially in the clinical world (e.g., electronic health records).

In the context of the Semantic Indexing of French Biomedical Data Resources (SIFR) project, we have developed the SIFR BioPortal (`http://bioportal.lirmm.fr`) [5], an open platform to host French biomedical ontologies and terminologies based on the technology developed by the US National Center for Biomedical Ontology [6,7]. The portal facilitates use and fostering of ontologies by offering a set of services such as search and browsing, mappings hosting and generation, metadata edition, versioning, visualization, recommendation, community feedback, etc. As of today, the portal contains 24 public ontologies and terminologies (+ 6 private ones) that cover multiple areas of biomedicine, such as the French versions of MeSH, MedDRA, ATC, ICD-10, or WHO-ART but also multilingual ontologies (for which only the French content is parsed) such as Rare Human Disease Ontology, OntoPneumo or Ontology of Nuclear Toxicity.

The SIFR BioPortal includes the SIFR Annotator[1] a publicly accessible and easily usable ontology-based annotation tool to process text data in French. This service is originally based on the NCBO Annotator [8], a web service allowing scientists to utilize available biomedical ontologies for annotating their datasets automatically, but was significantly enhanced and customized for French. The annotator service processes raw textual descriptions and tags them with relevant biomedical ontology concepts and returns the annotations to the users in several formats such as JSON-LD, RDF or BRAT. A preliminary evaluation [5] showed the web service matches previously reported work in French in performance, while being public, functional and turned toward the semantic web standards. However, this evaluation precedes the CLEF eHealth French task series and was not satisfactory. We had the motivation of evaluating the annotation service on the Quaero and CépiDC corpora used in the CLEF eHealth 2015-2017 French text data annotation tasks.

The SIFR BioPortal also offers the possibility of querying the original NCBO Annotator for English text via a dedicated proxy that extends the original

[1] `http://bioportal.lirmm.fr/annotator`

functionality. Thus, in this case, the +600 ontologies of the NCBO BioPortal[2] may be used. This service, called the NCBO Annotator+[3], is querying the original NCBO Annotator while offering new functionalities by pre processing of the input text and/or post processing of the original results. For instance, we have implemented the scoring of the results for both the SIFR and NCBO Annotator+ thanks to that proxy architecture [9]. Despite its wide and various uses and multiple evaluations, the NCBO Annotator has never been evaluated in the context of the CLEF eHealth tasks, and we believed it would be appropriate and relevant for the community to offer such an evaluation.

In this paper, we present our participation to the task 1 of the CLEF eHealth 2017 challenge, which tackles the problem of information extraction (diagnostic coding) in written clinical texts (death certificates). The objective of the task is to annotate each line of several death certificates, provided by the French Centre d'épidémiologie sur les causes médicales décès,(CépiDC) with an International Classification of Diseases, 10th revision (ICD-10) diagnostic code (French aligned task) or to annotate each document with the set of relevant ICD-10 diagnostic codes (French raw and English raw tasks). Considering that ICD-10 was never conceived to be used by automatic lexical tools, annotating the CépiDC data using only ICD-10 as source dictionary would have offered poor results, therefore, we have built custom SKOS vocabularies from the CépiDC/CDC dictionaries as well as the development and training corpora provided in the CLEF eHealth 2017 task 1 datasets. In the following, we will describe the construction of these custom vocabularies and present the results obtained both by the SIFR and NCBO Annotators used without any specific customization for the CLEF eHealth 2017 task 1. We obtained, for our best runs on each corpus, the following results: French Aligned (50.63% P, 52.97% R, 51.77% F1); French Raw (54.11% P, 48.00% R, 50,87% F1); English raw (69.08% P, 51.37% R, 58,92% F1). We will discuss the advantages and limitations of the annotators and possible perspectives for enhancing the performance of the specific task of coding death certificates or clinical notes. To us, in addition to technical performance (i.e., precision and recall) there are other aspects of the services that we think are crucial if we want to make the use of ontologies for annotation of clinical data mainstream. For instances, interoperability, ease of use as a service, openness, adoption of semantic web. A good annotation service shall be used in research or clinical environments, without any explicit knowledge of the technologies, the ontologies or the natural languages processing techniques involved.

2 Materials and methods

2.1 SIFR BioPortal and SIFR Annotator

The SIFR Annotator workflow is composed of several steps (figure to come): dictionary creation from ontologies, text pre-processing, concept recognition,

[2] `http://bioportal.bioontology.org`
[3] `http://bioportal.lirmm.fr/ncbo_annotatorplus`

semantic expansion (with mappings and hierarchy), annotations post-processing. For instance, in the final step, annotations are scored according to the context from which they have been generated, which is a requirement when they are used to index the original data. As another example, the SIFR Annotator can recognize negation, experiencer and temporality based on a customized French implementation of the NegEx/Context method [10]. Only the concept recognition step is evaluated in the context of the CLEF eHealth 2017 task 1, therefore, we will not describe into more detail the rest of the SIFR Annotator workflow here. For a presentation of the original NCBO Annotator service we point the readers to [8,11].

When calling the SIFR Annotator (Figure 1), the user's free text is given as input to a concept recognition tool along with a dictionary. The dictionary is a list of strings that identifies concepts, built out of any knowledge artifacts uploaded in the SIFR BioPortal (ontologies, terminologies, vocabularies, dictionaries). The dictionary is constructed by accessing these artifacts and pooling all string forms, such as preferred names and synonyms that syntactically identify concepts. The SIFR Annotator uses all the French ontologies and terminologies available in the SIFR BioPortal which give us a dictionary of 255K concepts and around twice that number of terms. Enabling the service to use another ontology is as simple as uploading the new ontology into the portal and the backend dictionary will be automatically regenerated.

The concept recognizer uses the complete dictionary for annotation, but users can filter out results to some specific ontologies depending of the type of biomedical data the service is used to annotate. The SIFR Annotator and NCBO Annotator rely on a fast string matching algorithm called Mgrep [12]. Mgrep is UTF-8 compliant (since v4) and can then be used with data in other languages than English. Mgrep and/or the NCBO Annotator have been evaluated multiple times [13,14,11,15,16] on different datasets an usually perform very well on precision e.g., 95% in recognizing disease names [17]. However, we do not know any specific evaluation of Mgrep performance on French. Mgrep can be parametrized (so can the SIFR Annotator) to match longest word only or all matches. It can also match partial part of the term in the dictionary. For a comparative evaluation of MetaMap [18] and Mgrep done when building the NCBO Annotator the reader may refer to [11]. Mgrep does not execute any natural language processing techniques such as lemmatization, stemming, spell-checking, or detection of morphological variants, but offers a fast and reliable (good precision) matching that enables use of the software in a web service environment. We therefore rely on the creation of the dictionary to augment the recall by adding alternate syntactic form or variants susceptible to be found in the text. For instance, the SIFR Annotator includes now a lemmatize mode (beta, not used within CLEF eHeath 2017 task 1) that will lemmatize the original input before sending it to Mgrep which then also uses a lemmatized dictionary.

Annotations performed with the SIFR Annotator can either be kept one by one with offset positioning of the matching term in the original given text, or can be aggregated (if multiple annotations with the same concepts occur in the

given text) and scored in order to be used for instance to index the original data [19]. An annotation (i.e., in essence, the URI of an identified concept) is always described with contextual information describing whether the annotation is direct or semantically expanded, if the match was done on a preferred name or synonym or if the identified concept is for instance negated. Neither scoring nor context information aspects are used in the CLEF eHealth 2017 task 1 settings.

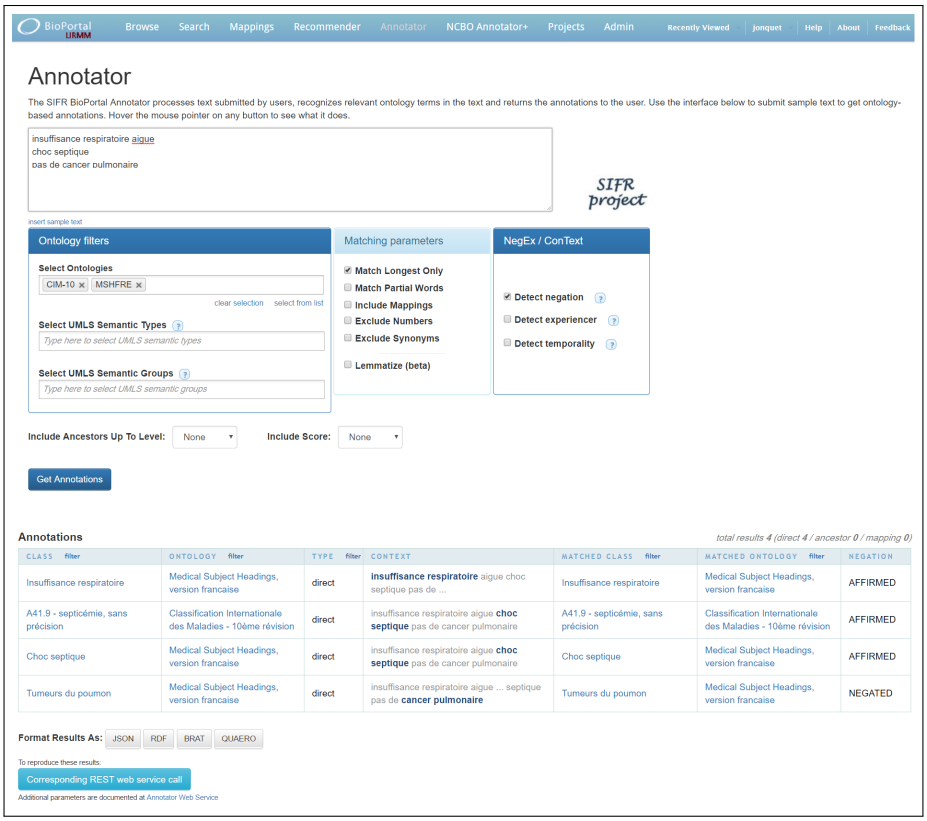

Fig. 1: Annotator interface in the SIFR BioPortal. The text is here annotated with Mesh and CIM-10 (ICD10), with negation and longest only parameters.

An important aspect for the SIFR Annotator is to be available as a web service. The service results may be described in multiple syntax (XML or JSON-LD) and format (e.g., RDF/XML described with the Annotation Ontology or BRAT). A specific CLEF eHealth output format has been also implemented to evaluate the service against the previous campaigns. For testing and ontology selection purposes, the web application user interface is available (Figure 1), however, the service is meant to be used through the REST application programming interface.

Note that thanks to a Docker (`www.docker.com`) packaging the SIFR BioPortal and Annotator may easily be installed locally to process sensitive data in-house, a common requirement when manipulating clinical data. All the code is open source and available on GitHub (`https://github.com/sifrproject`).

Within the SIFR project, we also developed an enhanced version of the NCBO Annotator to annotate English biomedical text data, without having to serve English ontologies locally. The NCBO Annotator+ uses a proxy service architecture that enhances the capabilities of the original annotation service by encapsulating around the original application programming interface. For instance, our recent implementations done for the French annotator are available, thanks to that proxy architecture, also for English e.g., scoring of annotations [9] or more recently detection of negation.

For CLEF eHealth 2017 task 1, we have used the SIFR and NCBO Annotators "as it is" without any specific customization for the task nor any training on the data. The runs have been executed with longest only parameter on, and without semantic expansion of the annotations, scoring or contextualization.

2.2 Task and corpus

The objective of CLEF eHealth 2017 task 1 is to annotate death certificates with ICD-10 codes both in French and in American English. For English, a corpus of death certificates from the CDC was provided, separated in a training and a test corpus. The training corpus contains 13,329 death certificates, for a total of 32,714 lines. The test corpus contains 6,665 certificates containing a total of 14,834 lines. For French, a corpus of death certificates from CépiDC was provided. More specifically, there was a training corpus that contained 65,844 documents and 195,204 lines, a development corpus with 27,851 document and 80,900 lines and a test corpus with 31,683 documents and 91,954 lines. The corpora are digitized versions of actual death certificates filled in by clinicians. Although the punctuation is not always correct or present, the corpus is already segmented in lines (as per the standard international death certificate model) which for the most part contains single sentences.

The French corpus was provided in both an aligned and a raw format, while the English corpus was only provided in the raw format. The raw format provides two files, a `CausesBrutes` file and an `Ident` file. The former contains semicolon separated values for the Document identifier (DocID), the year the certificated was coded (YearCoded), the line identifier (LineID), the raw text as it appears in the certificate (RawText), an interval type during which the condition occurred (IntType - seconds, minutes, hours, weeks, years) and an interval value (IntValue). The `Ident` file contains a document identifier, the year the certificate was coded, the gender of the person, the code for the primary cause of death, the age and the location of death. The aligned format is a reconciliation between the `CausesBrutes` and `Ident` files, where the fields have been aligned at the document and line number level. Thus, the aligned file contains the same unique fields that the original ones to which an extra field is added in the gold standard dataset providing a standardized text that represents the manually annotated code.

For CLEF eHealth 2017 task 1, we have used only the "RawText" information of both the aligned and raw datasets. We did not use any other information/features such as age or gender contained in the files.

2.3 Dictionaries construction

SIFR BioPortal already contained the French ICD-10[4] (CIM-10) reference terminology. This OWL version was originally produced by the CISMeF team from an automatic export from the HeTOP ontology/terminology server [20]. Respectively, the NCBO BioPortal already contained the English ICD-10.[5] This RDF version was automatically exported from the Unified Medical Language System (UMLS) with the umls2rdf tool. [6] However, the purpose of ICD-10 is to serve as a general purpose reference to code medical acts, and not to be directly used for text annotation and, especially not in a particular clinical task such as death certificate coding. Indeed, from our experiments, using ICD-10 classification alone for annotation leads to a F1 score below 10%.

For the French tasks, a set of dictionaries was provided by CépiDC that give a standardized description text of each of the codes that appear in the corpora. Additionally, the data from the aligned corpus (French only) could also be used to enrich the lexical terms of ICD-10. A similar dictionary was provided for the English task. In order to use these dictionaries within the SIFR and NCBO Annotator, we had to encode them using a format accepted as input within the portal, which includes RDFS, OWL, SKOS, OBO or RRF (UMLS format). In this case, the ideal choice in terms of standardization, potential reusability and simplicity was to use SKOS (Simple Knowledge Organization System) a W3C Recommendation specialized for vocabularies and thesaurus. For CLEF eHealth 2017 task 1, we produced two groups of SKOS dictionaries: CIM-10DC* for French, based on the French dictionaries and aligned corpus; ICD-10-CDC* for English based on the CDC corpus dictionary alone.

We set out in this construction process by first defining the appropriate schema to represent the SKOS dictionaries. We chose to use the same URIs as concepts identifiers for `skos:Concept` that the `owl:Class` in the available CIM-10/ICD-10 which allows our dictionaries to be fully aligned ontologically speaking with the original terminologies they enrich. Each of the codes was represented by a `skos:Concept`. The URIs are composed of a base URI and a class identifier that represents the ICD codes, in the following format: "[A-Z][0-9][0-9]?[0-9]?" (e.g. G12.1 or A10). The identifier is slightly different from the codes from the task dictionaries: there is a dot before the last digit and if the last digit is zero, then the dot and the last zero are omitted. Thus, G12.1 in CIM-10 corresponds to G121 in the corpus while A10 in CIM-10 corresponds to A100 in the corpus. The corresponding URI in CIM-10 and this in CIM-10DC are: `http://chu-rouen.fr/cismef/CIM-10#G12.1`, where `http://chu-rouen.fr/cismef/CIM-10#` is the

[4] `http://bioportal.lirmm.fr/ontologies/CIM-10`
[5] `https://bioportal.bioontology.org/ontologies/ICD-10`
[6] `https://github.com/ncbo/umls2rdf`

base URI and G12.1 the code identifier. In ICD-10 and ICD-10CDC the URIs are like so: `http://purl.bioontology.org/ontology/ICD-10/P08.0`, where the base URI is `http://purl.bioontology.org/ontology/ICD-10/` and the code identifier is P080.

When building the SKOS dictionaries, we used the same chapter hierarchy as ICD-10, although we kept only the chapter level of the hierarchy. This was mainly done for convenient browsing and visualization of the dictionaries in the NCBO and SIFR BioPortal. Each code concept was attached below the corresponding chapter concept (e.g. I98.5 belongs to the chapter I9.00) with the `skos:broader` property.

Construction algorithm We built the French SKOS dictionary from the aligned corpus and all the CépiDC dictionaries. We built the English one only from the raw corpus and the CDC dictionary. We first built a code index, that to each code associated the list of labels retrieved from:

- The `DiagnosisText` field in the dictionary, associated to codes through the ICD1 and ICD2 fields. For French, we used a concatenation of all the dictionaries and for English we used the one dictionary file provided.
- The `RawText` and `StandardText` (only for French) fields from the corpus associated to codes through the ICD-10 field in the corpus (AlignedCauses_2013Full for French, CausesCalculees_EN_training for English).

For each code concept the CépiDC and CDC dictionaries contained multiple labels. In order to follow SKOS specification, we had to select a preferred name automatically (`skos:prefLabel`) and assign the other labels as alternative labels (`skos:altLabel`). Note that this selection would note influence the annotation process as both preferred name and synonyms are included in the concept recognizer dictionaries. The selection heuristic took the shortest label that does not contain three or more consecutive capital letter (likely and acronym).

Ambiguous label selection heuristics An important issue when building the SKOS dictionaries was to assign ambiguous labels (i.e., identical labels which correspond to different codes). Indeed, those labels create ambiguity in the annotations and leads to better recall at the price of a low precision. For example, the label "choc septique" was present as preferred label or synonyms for 58 different codes. Therefore, we had to implement a selection heuristic to determine the most suitable code to which the label should be bound.

When using both the standard text and the raw text fields from the corpus, if standard text label is ambiguous, a simple heuristic is to not add it to any code but just use the raw text instead. Given that the raw text is unique, the ambiguity related to the inclusion of the test corpus is removed. We called this first strategy "Adaptive dictionary generation" and created a CIM-10DCA French SKOS dictionary to evaluate it. Given that this strategy relied on the availability of a standardized text, it was confined to the French aligned task, as the raw English corpus contained no standard labels.

The drawback with the previous heuristic is that we lose some labels that would otherwise have potentially increased recall. Thus, we searched a way of assigning ambiguous labels to one code only. Taking inspiration from the idea of the most frequent sense baseline often used in Word Sense Disambiguation tasks, we adopted a heuristic that assigns ambiguous labels to the most frequent code only. We use the training corpus to estimate the frequencies of use of the codes (gold standard annotations) so that when a label can belong to several codes, we can sort the codes by frequency and chose either the most frequent code (MFC) or the top k most frequent codes (kMFC).

In practice the "Adaptive" strategy led to a much lower recall without particularly improving precision. Final F1 scores were worse than with the MFC strategy which led to a precision and recall that were balanced. This is the strategy we have finally used in the reported results.

Availability of the SKOS dictionaries The final SKOS dictionaries built with the best ambiguous label strategy have been uploaded respectively on the SIFR and NCBO BioPortal, and are accessible in private mode only for now (access upon request, an account is required):

- CIM-10DC-ALL contains 6817 concepts for a total of 295,385 labels.
- CIM-10DC-ALLMFC contains the same number of concepts but only 249,524 labels.
- ICD-10CDC contains 3738 concepts with 166,500 labels.

We have also created a resource solely from the CépiDC dictionaries (without using the corpus) called CIM-10DCD. This one do not contain any sentence originally present in death certificates. We are currently discussing with the CépiDC, potential public releases of the SKOS dictionaries as well as the workflow to update them on a regular basis. Indeed, we believe it is also part of the SIFR project mission to facilitate open access to resources (and adopting standard ways of describing them e.g., semantic web standards), when licensing permits.

2.4 ICD-10/CIM-10 Coding with ontology concepts

Given that we used the SIFR and NCBO Annotators, besides manually curating the created SKOS dictionaries, the final step to obtaining a working system for the task was to write a complete workflow to:[7]

1. Read the corpus in the raw or aligned formats;
2. Send the text to the Annotators REST API with the right ontologies and annotation parameters;
3. Retrieve the annotations produced by the Annotators;
4. Optionally prune some annotations (post-annotation heuristic);
5. Produce the output in the right raw or aligned format.

[7] For this purpose, we used the Java language.

We implemented two post-annotation heuristics:

- Most Frequent Code, if a particular line was annotated with several codes, only keep the most frequent code based on the code distribution of the training corpus. There are no parameters.
- Code Frequency Cutoff, we calculate a normalized probability distribution of the codes that annotate a particular line (with a density function in the [0, 1] range) and only keep the codes below a cumulative probability threshold. The only parameter is the cut-off threshold.

The parameters of the entire system are the combination of the parameters of NCBO or SIFR annotator (list of ontologies, longest_match (T/F), expand_mappings (T/F)) with any post-annotation heuristic parameter.

2.5 Reproducibility

Using the SIFR annotator is as simple as sending a request through the HTTP REST API, for example to annotate the sentence: "Absence de tumeur maligne" with the French version of WHO-ART and MedDRA, it can be done with:[8]

`http://services.bioportal.lirmm.fr/annotator/?text=Absence%20de%20tumeur%20maligne&negation=true&ontologies=WHO-ARTFRE,MDRFRE`

Because of the sensitive nature of the CépiDC data, we have used a local version of the NCBO and NCBO Annotators in order to avoid sending the data out on the network. However, to reproduce our results one would require to have access to the private SKOS dictionaries on the NCBO and SIFR BioPortal and in that case, use the user interface or the REST web service API. The reproduction instructions are available here: `https://twktheainur.github.io/bpannotatoreval/LIRMMCLEF2017Task1Instructions.html`

3 Results

Using the previously described workflow, we have performed six runs:

French Run 1 (Aligned and Raw): Annotation with longest only parameter on and running on a local instance of the SIFR Annotator with CIM-10 and CIM-10DC-ALLMFC as target resources.

French Run 2 (Aligned and Raw): A fallback strategy starting from the result file of Run 1, and, for each line without any annotations, takes the annotations from a second run, which used CIM-10 and CIM-10DC-ALL as target resources. This is, in essence a late fusion technique, that aims at increasing the recall, without sacrificing precision.

[8] The REST API requires an APIkey to be used (obtained by creating an account on the portal). In that example, one can use the demo API key and add &apikey=c34b7653-0639-4946-81af-8ac76fe809dd at the end of the call.

English Run 1 (Raw): Annotation with longest only parameter on and running on a local instance of NCBO Annotator with ICD-10 and ICD-10CDC as target resources.

English Run 2 (Raw): Same as Run 1 but with ICD-10CM as additional target (also available in the NCBO BioPortal): the Clinical Modification of ICD-10 made in the USA for the classification of morbidity causes.

3.1 English raw results

11 teams participated to the English raw evaluation. Table 1 presents the results obtained by our two runs against the average and median results of the runs submitted to this task. The NCBO Annotator obtained results that are exactly the median value of all the results submitted (all causes). We can measure a slight decrease in precision and increase in recall with the introduction of ICD-10-CM in Run 2. Regarding the external causes, the NCBO Annotator obtains a better precision and f-measure than the average and median results submitted to the challenge.

Table 1: Results on the English raw dataset

	All Causes			**External Causes**		
	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
Run1	69.1	51.4	58.9	23.2	52.4	32.2
Run2	64.6	52.7	58.0	23.3	52.4	32.3
Average	67.0	58.2	62.2	40.5	26.7	26.1
Median	64.6	60.6	61.1	27.9	26.2	27.4

3.2 French raw and aligned results

13 runs have been submitted by 9 teams to the French raw evaluation. 7 runs have been submitted by 5 teams to the French aligned evaluation. Tables 2 and 3 present the results obtained by our two runs against the average and median results of the runs submitted to this task. As expected, the SIFR Annotator did perform similarly on the raw and aligned datasets (as they were processed exactly with the same workflow). The results are exactly the median value of all the results with the raw dataset, but slightly under the median value for the aligned datasets (all causes). Indeed, teams that have used other information from the aligned dataset probably performed better than the SIFR Annotator here. Regarding the external causes, we obtain better precision and F1 than the average and median results submitted to the challenge.

Table 2: Results on the French raw dataset

	All Causes			External Causes		
	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
Run1	54.1	48.0	50.9	44.3	36.7	40.1
Run2	54.0	48.0	50.8	44.3	36.7	40.1
Average	47.5	35.8	40.6	36.7	24.7	29.2
Median	54.1	41.4	50.8	44.3	28.3	37.6

Table 3: Results on the French aligned dataset

	All Causes			External Causes		
	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
Run1	50.6	53.0	51.8	41.2	40.3	40.7
Run2	50.5	53.0	51.7	41.2	40.3	40.7
Average	64.8	55.6	59.3	50.5	31.9	36.6
Median	62.9	54.0	54.8	50.8	33.3	40.6

4 Discussion

The results obtained are in line with what could be expected from our approach, which really is an evaluation of how the SIFR and NCBO Annotators concept recognition component works. The simple string matching approach adopted by Mgrep generally offers a good precision (around 80%) if the ontology is properly lexicalized and concretely captures the terms used in the text to annotate. For CLEF eHealth 2017 task 1, the loss in precision is explained by the nature of the dictionaries used as a source to produce our terminology: a same label can correspond to several different classes (here ICD10 codes). This practice is usually strongly avoided when designing ontologies as it inevitably creates ambiguities. Concerning recall, our performance is also limited since the concept recognizer does not include any natural language processing techniques that would increase the amount of matches, handle morphological variants (as simple as plural forms) or any other alternative concepts. All phenomena that are common in reality but not captured as synonyms by the source ontologies will not be recognized properly. Previous evaluation of the NCBO Annotator [11] already identified such limitations. Unsurprisingly, we found the same limitations apply to the French version. These issues are particularly important when using resources such as ICD10 that are not designed to be automatically used for annotation (in spite of it's original mission that is indeed coding medical acts).

The purpose of the SIFR Annotator, as the NCBO Annotator when released in 2009, was not to beat state-of-the-art systems tuned for specific tasks. The concrete advantages of the services, both respectively connected to the SIFR and NCBO BioPortal come from: (i) the size and variety of their dictionaries coming from ontologies, (ii) their availability as a web service that can be easily connected in any semantic indexing workflow, and finally (iii) their adoption of a semantic web vision that strongly encourages returning URIs that can then reused to facilitate data integration and semantic interoperability. One should also note that the semantic expansion step (which uses the mappings between ontologies and the is_a hierarchies to generate additional annotations) as well as the post-processing of the annotations (which scores and contextualizes the annotations) are interesting features not evaluated within CLEF eHealth 2017 task 1.

Despite of their limitations, the NCBO and SIFR Annotators obtained exactly median results when compared to the performance of all the participating systems. Therefore, considering the other discussed advantages, we believe they are two services that can help in a wide class of text mining or annotation problems, but of course not for all. It is important to note the systems were not tailored for this task and their performance will highly vary depending of the data to annotate and the ontologies targeted.

Participation in CLEF eHealth 2017 task 1 is, however, a very good manner of enabling us to improve our SIFR Annotator and potentially the NCBO Annotator also. Such improvements shall be either generic (changes to the overall workflow, independent of task) or tailored for improving the results to the CLEF eHealth series (Quaero or CépiDC corpora). In order to better understand the shortcoming of the system, we sampled 100 false positives and false negatives from the best runs of the SIFR Annotator on the French aligned development dataset and proceeded to manually determine the case of the error. Some class of errors are as follows:

- Errors because of missing synonyms. A few good illustrative examples include:
  - The code R09.2 “arrêt respiratoire” was not identified within the text “arrêt cardio respiratoire” or “détresse cardiorespiratoire.”
  - The code J96.0 “insuffisance respiratoire aiguë” was not identified within the text “détresse respiratoire.”

  Although some of these false negative could be avoided with a richer dictionary (1st case) or simple synonym generation (2nd case), we found some that strongly relies on some medical expertise that can hardly be captured by a dictionary based approach (maybe by a machine learning one, assuming there is enough data to train the tool).
- Single match returned whereas a multiple match was expected. Indeed, the MFC strategy resulted in assigning the synonyms to only one code in CIM-10-DC therefore, when for the same text, several annotations where expected, we found only one. For instance, the code G40.9 ”épilepsie, sans précision” was found with the text ”epilepsie avec etat de mal” but not the code G41.9 ”état de mal épileptique, sans précision.” Note that those errors are only present with the MFC strategy.

- Morphosyntactic or lexical variation (e.g., accent, dash, comma, spelling). For instance, the code J18.9 “emphysème, sans précision” was not identified within the text “emphysème pumonaire secodaire tabagisme actif” because of the spelling of ”pulmonaire.”
- Annotations were made with a more general (i.e., parent in ICD-10 hierarchy), often because of a partial match within an expression.
- Errors because the annotation would require to use other explicit information in the certificate or medical knowledge. For instance, the annotation with code I10 “hypertension essentielle (primitive)” was not found from the text “TC suite à une chute avec épilepsie séquellaire et tr cognitifs” as it was annotated within the corpus. Or the code R68.8 “autres symptômes et signes généraux précisés” was not identified within the text “atteinte polyviscérale diffuse.”

From this review of the pitfalls of the SIFR Annotator on the CépiDC corpus, and from other in-house experiments, we clearly identified the need to improve the dictionary generation process when extracting the labels from the source ontologies. This is the moment where the terms shall be enriched adding other alternative synonyms or morphosyntactic or lexical variations. However, this shall be done with precautions to be sure to augment recall without decreasing precision as the introduction of new ambiguity shall be avoided.

Another potential important source of improvement for the SIFR Annotator shall come from generating and curating alignments between ontologies. For instance, on the CépiDC corpus, when used with other ontologies the SIFR Annotator was able to identify unambiguous concepts from other ontologies. The system is implemented to use the mappings between ontologies to expand the original direct annotations made from the text, but of course this suppose the mappings have been generated and uploaded to the SIFR BioPortal. In the case of ICD10, there exists multiple source of published mappings that we plan to upload in the future. This will clearly have an effect on recall.

The limitations on one annotation scenario shall not be the same on another scenario. We are currently evaluating the performance of the SIFR Annotator on the Quaero corpus [21] used during previous editions of CLEF eHealth and we identified other problems such as ambiguity (as we use several ontologies whereas Quaero is annotated with unique UMLS Metathesaurus concepts) or missing translated terms (as the Quaero corpus used English UMLS concept when French ones were not available thus inevitably disadvantaging methods that would not use some kind of translation approach). Overall, despite of these limitations, our results on the Quaero corpus varies between 63 and 70% F1 on plain entity recognition (UMLS semantic groups) and 36-37% F1 on normalized entity recognition (UMLS concept unique identifiers) on the EMEA dataset. And respectively 58-69% F1 and 32-33% the Medline dataset. These results put the SIFR Annotator among the top systems for plain entity and in the first half for normalized entity (while not using automatic translation!). Better Quaero results analysis and reporting shall be the subject of another specific future communication.

We would also like to point to some recent improvements that we have made to the SIFR Annotator that are currently under evaluation. When annotating clinical notes with medical conditions, it is important to filter out negated conditions and distinguish present conditions from antecedents or conditions experienced by someone other than the patient. For this reason, several methods have been proposed to detect the context of already identified clinical conditions, especially for English medical text. The English-language system, NegEx/ConText, is one of the best and fastest algorithms for the determination of the context of medical conditions [10]. ConText is based on lexical cues (trigger terms) that modify the status of medical conditions appearing in the scope of the cues. We have adapted this system to the French language. Our approach consisted in compiling an extensive list of French lexical cues by a process of automatic and manual translation and enrichment. Then, we interconnected the NegEx/ConText program with the NCBO and SIFR Annotators thanks to the proxy architecture previously mentioned. This feature can already be used on the SIFR BioPortal for the French and English annotation services. Our first evaluation confirms the ability to detect negation with a very high F1 score, slightly improving previous published work done in the past to adapt NegEx for French. This study shall also be part of another specific future communication on the SIFR Annotator. Due to time limitation, we have not use NegEx/ConText on the CépiDC corpus although we are not sure about the impact of this feature on the results.

5 Conclusions

In this paper, we presented our participation to the task 1 of the CLEF eHealth 2017 challenge using the NCBO and SIFR Annotators. Our results are encouraging: around 50-60% of F1 score means that more than half of the task of coding death certificates with ICD-10 codes can be automatized. Especially considering that we have not implemented anything specific to process these data. But of course, we will have to improve these results to be among the best performing systems. Some improvements perspectives have been discussed.

We have also argued in this paper that according to us the technical performance (F1) shall not be the only argument in evaluating a semantic annotation tool. The SIFR project has invested a significant amount of effort in order to offer a generic, open and quite robust platform that could easily be used "at the click of the mouse" to annotate biomedical text data and access French biomedical ontologies and terminologies. Someone can upload a new resource to the SIFR BioPortal and get a dedicated annotation service, interconnected to other existing ontologies, in a couple of hours.

Indeed, the SIFR Annotator is different from other related tool in French biomedical text mining as: (i) it is a dynamic web service with JSON-LD outputs which can be integrated in current programmatic workflows; (ii) it uses public ontologies both to create annotations and to expand them; (iii) it has access to one of the largest available sets of publicly available biomedical ontologies in French. We believe the SIFR Annotator can therefore be used in a large

span of biomedical applications including annotating clinical text data. We are currently using the service in the context of the French PractiKPharma project (`http://practikpharma.loria.fr`) which aims to validate pharmacogenomics state-of-the-art knowledge on the basis of practice-based evidences, i.e., knowledge extracted from electronic health records.

Acknowledgements

This work is supported by the French National Research Agency within the PractiKPharma (grant ANR-15-CE23-0028) and SIFR (grant ANR-12-JS02-01001) projects as well as by the European H2020 Marie Curie actions (grant 701771), the University of Montpellier and the CNRS. We also thanks the US National Center for Biomedical Ontology for their assistance with the NCBO Annotator and the CLEF eHealth 2017 organizers for their help.

References

1. Blake, J.A.: Bio-ontologies—fast and furious. Nature Biotechnology **22** (June 2004) 773–774
2. Rubin, D.L., Shah, N.H., Noy, N.F.: Biomedical ontologies: a functional perspective. Briefings in Bioinformatics **9**(1) (2008) 75–90
3. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F.: Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semantics: Science, Services and Agents on the World Wide Web **4**(1) (January 2006) 14–28
4. Névéol, A., Grosjean, J., Darmoni, S.J., Zweigenbaum, P.: Language Resources for French in the Biomedical Domain. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: 9th International Conference on Language Resources and Evaluation, LREC'14, Reykjavik, Iceland, European Language Resources Association (May 2014) 2146–2151
5. Jonquet, C., Annane, A., Bouarech, K., Emonet, V., Melzi, S.: SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In: 16th Journées Francophones d'Informatique Médicale, JFIM'16, Genève, Suisse (July 2016) 16
6. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N.B., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research **37**((web server)) (May 2009) 170–173
7. Whetzel, P.L., Team, N.: NCBO Technology: Powering semantically aware applications. Biomedical Semantics **4S1**(S8) (April 2013) 49
8. Jonquet, C., Shah, N.H., Musen, M.A.: The Open Biomedical Annotator. In: American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09, San Francisco, CA, USA (March 2009) 56–60
9. Melzi, S., Jonquet, C.: Scoring semantic annotations returned by the NCBO Annotator. In Paschke, A., Burger, A., Romano, P., Marshall, M., Splendiani, A., eds.: 7th International Semantic Web Applications and Tools for Life Sciences,

SWAT4LS'14. Volume 1320 of CEUR Workshop Proceedings., Berlin, Germany, CEUR-WS.org (December 2014) 15
10. Harkema, H., Dowling, J.N., Thornblade, T., Chapman, W.W.: Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of biomedical informatics **42**(5) (2009) 839–851
11. Shah, N.H., Bhatia, N., Jonquet, C., Rubin, D.L., Chiang, A.P., Musen, M.A.: Comparison of concept recognizers for building the Open Biomedical Annotator. BMC Bioinformatics **10**(9:S14) (September 2009)
12. Dai, M., Shah, N.H., Xuan, W., Musen, M.A., Watson, S.J., Athey, B.D., Meng, F.: An Efficient Solution for Mapping Free Text to Ontology Terms. In: AMIA Symposium on Translational BioInformatics, AMIA-TBI'08, San Francisco, CA, USA (March 2008)
13. Simon N. Twigger, Joey Geiger, J.S.: Using the NCBO Web Services for Concept Recognition and Ontology Annotation of Expression Datasets. In Marshall, M.S., Burger, A., Romano, P., Paschke, A., Splendiani, A., eds.: Workshop on Semantic Web Applications and Tools for Life Sciences, SWAT4LS'09. Volume 559 of CEUR Workshop Proceedings., Amsterdam, The Netherlands, CEUR-WS.org (November 2009)
14. Sarkar, I.N.: Leveraging Biomedical Ontologies and Annotation Services to Organize Microbiome Data from Mammalian Hosts. In: American Medical Informatics Association Annual Symposium, AMIA'10, Washington DC., USA (November 2010) 717–721
15. Groza, T., Oellrich, A., Collier, N.: Using silver and semi-gold standard corpora to compare open named entity recognisers. In: Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on, IEEE (2013) 481–485
16. Funk, C., Baumgartner, W., Garcia, B., Roeder, C., Bada, M., Cohen, K.B., Hunter, L.E., Verspoor, K.: Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. BMC bioinformatics **15**(1) (2014) 59
17. Xuan, W., Dai, M., Mirel, B., Athey, B., Watson, S.J., Meng, F.: Interactive Medline Search Engine Utilizing Biomedical Concepts and Data Integration. In: BioLINK: Linking Literature, Information and Knowledge for Biology, SIG, ISMB'08, Vienna, Austria (July 2007) 55–58
18. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: American Medical Informatics Association Annual Symposium, AMIA'01, Washington, DC, USA (November 2001) 17–21
19. Jonquet, C., LePendu, P., Falconer, S., Coulet, A., Noy, N.F., Musen, M.A., Shah, N.H.: NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources. Web Semantics **9**(3) (September 2011) 316–324
20. Grosjean, J., Merabti, T., Griffon, N., Dahamna, B., Darmoni, S.: Multiterminology cross-lingual model to create the European Health Terminology/Ontology Portal. In: 9th International Conference on Terminology and Artificial Intelligence, TIA'11, Paris, France (November 2011) 119–122
21. Névéol, A., Grouin, C., Leixa, J., Rosset, S., Zweigenbaum, P.: The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In: Proc of BioTextMining Work. (2014) 24–30