Enrichment of French Biomedical Ontologies with UMLS Concepts and Semantic Types for Biomedical Named Entity Recognition Though Ontological Semantic Annotation.

Andon Tchechmedjiev
Université de Montpellier, LIRMM
andon.tchechmedjiev@lirmm.fr

Clément Jonquet
Université de Monpellier, LIRMM
Center for Biomedical Informatics Research,
Stanford University
jonquet@lirmm.fr

September 6, 2017

Abstract

Medical terminologies and ontologies are a crucial resource for semantic annotation of biomedical text. In French, there are considerably less resources and tools to use them than in English. Some terminologies from the Unified Medical Language System have been translated but often the identifiers used in the UMLS Metathesaurus, that make its huge integrated value, have been lost during the process. In this work, we present our method and results in enriching seven French versions of UMLS sources with UMLS Concept Unique Identifiers and Semantic Types based on information extracted from class labels, multilingual translation mappings and codes. We then measure the impact of the enrichment through the application of the SIFR Annotator, a service to identify ontology concepts in free text deployed within the SIFR BioPortal, a repository for French biomedical ontologies and terminologies. We use the Quaero Corpus to evaluate.

1 Introduction

As of early 2017, the Linked Open Data cloud diagram[1] became largely dominated by life-sciences and more specifically, by biomedical ontologies and terminologies hosted on the BioPortal repository developed by the US National Center for Biomedical Ontology (Noy et al., 2009). The NCBO BioPortal, is a reference ontology repository for the biomedical domain that provides open and accessible ontology indexing, browsing, search recommendation and semantic annotation. NCBO BioPortal includes, as of Summer 2017, more than 580 language resources, but only few are not in English, e.g., five in French and one in Spanish (Jonquet et al., 2015). Furthermore, the UMLS (Unified Medical Language System) Metathesaurus (Bodenreider, 2004), even if it covers 21 languages, 75.1% of its terms are in English and only 1.82% of its terms are in French (Bollegala et al., 2015).

Our work is part of the SIFR project (Semantic Indexing of French Biomedical Data Resources - http://www.lirmm.fr/sifr) in which we are interested in exploiting ontologies in construction of services like indexing, mining, and information retrieval for French biomedical resources. In this project, we develop a semantic indexing workflow (called the French/SIFR Annotator) based on ontologies similar to that existing for English resources [16], but focused on the French resources. The present study concerns 7 French terminologies hosted on the SIFR BioPortal (http://BioPortal.lirmm.fr) (a local instance of

[1] http://lod-cloud.net

BioPortal dedicated to French) that we wished to formally enrich with UMLS concepts and semantic type identifiers.

To improve the SIFR Annotator workflow and enable the use of UMLS identifiers, we present our method and results in enriching seven French medical terminologies with UMLS Concept Unique Identifiers (CUIs) and Semantic Type identifiers (TUIs). The English version of the seven processed terminologies are included within the UMLS Metathesaurus, but the original concept and type identifiers have not been ported to their French version, when translated. This was a big limitation for users interested in manipulating the French version of the terminologies while leveraging the manual original semantic integration effort made when the English version were included in the Metathesaurus.

The lack of anchorage of translated medical terminologies in the UMLS represents a real barrier for non-English-speaking communities that produce and manage biomedical data in their own languages. For example, France, Spain, Italy or Germany. UMLS concepts and semantic types are often used as gold standard annotations is most annotation tasks/campaigns for biomedical information extraction (e.g. some tasks of the CLEF eHealth evaluation campaign in 2015 and 2016 with the Quaero corpus (Névéol et al., 2014)).

To ensure semantic interoperability it is not enough to just translate ontologies, we must also formally keep the link between objects of the translated ontologies and the original one. Such data also needs to be semantically represented to be exploitable by machines (e.g., Linked Open Data vision). In previous work, we have reconciled more than 228K mappings between ten English ontologies hosted on NCBO BioPortal and their French translations hosted on the SIFR BioPortal. But still, the UMLS identifiers were missing. Re-establishing the broken links between English UMLS sources and their French counterpart, not included in the UMLS, was the aim of this work.

In the remainder of the paper, we first present background and related work about French medical terminologies and their relation to UMLS. Subsequently, we present the enrichment methodology and algorithm based on information extracted from class labels, multilingual translation mappings and codes. Then we evaluate the impact of the enrichment on the SIFR Annotator performance on the Quaero corpus, before concluding and giving some future perspectives.

2 Related Work

2.1 SIFR BioPortal

In the context of the Semantic Indexing of French Biomedical Data Resources (SIFR) project, we have developed the SIFR BioPortal (`http://BioPortal.lirmm.fr`) Jonquet et al. (2016), an open platform to host French biomedical ontologies and terminologies based on the technology developed by the US National Center for Biomedical Ontology (Noy et al., 2009; Whetzel and Team, 2013). The portal facilitates the use and fostering of ontologies by offering a set of services such as search and browsing, mapping 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[2] 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 (Jonquet et al., 2009), 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 input by users, 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 Jonquet et al. (2016) showed that the web service matches the results of previously reported work in French, while being public, functional and turned toward semantic web standards. SIFR

[2] `http://BioPortal.lirmm.fr/annotator`

Annotator allows users to input free text and to annotate the text with ontology concepts. SIFR Annotator, uses a dictionary composed of a flat list of terms build the concept labels and synonym labels from all the resources uploaded in SIFR BioPortal (ontologies, terminologies, vocabularies, dictionaries). SIFR BioPortal currently contains about 255K concepts and around twice that number of terms.

Enabling the service to use additional ontologies is as simple as uploading them to the portal (the indexing and dictionary generation are automatic).

2.2 Ontology Alignment and French Biomedical Ontologies

There have been initiatives in the past to reinforce the involvement of French language in the UMLS which contains now 5 French terminologies (Darmoni et al., 2003; Zweigenbaum et al., 2003; Annane et al., 2016). However, most of the French ontologies and terminologies are still not included; they are most often aggregated and translated by the CISMeF group[3] (Grosjean et al., 2011) (324.000 French concepts in HeTOP vs. 85,000 in the native UMLS) and since more recently also offered within the SIFR BioPortal (Jonquet et al., 2016).

There are very few attempts at aligning French biomedical terminologies/ontologies between each other or with equivalent English-language ontologies. The UMLS Metathesaurus itself can be considered as a large scale ontology alignment initiative, as it constitutes a pivot-based alignment of medical terminologies in several languages (Bodenreider et al., 1998). As for French-specific ontology translation and alignment, the work on MeSH by the French organization INSERM[4] is a good example. However, the most important effort in France is achieved by the Rouen University Hospital within the context of the CisMeF project (Merabti et al., 2012).

When integrating and translating new terminologies within the HeTOP platform(Grosjean et al., 2011) , they performed they generally aligned the new content with the UMLS. Although that information was poorly represented (e.g., CUIs were encoded as labels)s in the OWL version exported from HeTOP and imported into the SIFR BioPortal, we reused that information during our enrichment process. .

Previous work by Annane et al. (2016) explored the reconciliation of the French terminologies and ontologies in the SIFR BioPortal with their equivalent English ontologies within the NCBO BioPortal. Now, the locally hosted ontologies are formally aligned and the alignments are available within the SIFR BioPortal, adapted to allow interportal mappings. In most cases, the mappings were produced through a code reconciliation between the ontologies. We have used these multilingual translation mappings in the present work.

Even in English, there is little work related to enriching existing English-language biomedical ontologies with UMLS CUIs, let alone French-language ontologies.Rajput and Gurulingappa (2013) use direct concept name matching to establish a correspondence between UMLS and their own neurodegenerative disease ontology composed of 1147 concepts. Sarkar et al. (2003) apply a range of ontology matching techniques (exact-match, match on normalized UMLS strings and using MetaMap) to enrich the Gene Ontology (GO) with UMLS semantic information. There are, to our knowledge no attempts at enriching French biomedical ontologies and terminologies automatically with UMLS concepts and semantic types.

The UMLS group within the SIFR BioPortal contains 10 medical terminologies (Table 1). Three terminologies (highlighted in gray in Table 1) where directly extracted from the UMLS with a customized version of the NCBO developed umls2rdf tool (`https://github.com/sifrproject/umls2rdf`). For these three terminologies no enrichment was necessary, as the output generated by the tool already included UMLS CUIs and TUIs. The rest of the seven ontologies (highlighted in blue in Table 1) where generated by an OWL export from the HeTOP platform and although they English counterpart was included in the UMLS, the French version did not have CUI and TUI information.

[3]Rouens University Hospital (`http://www.chu-rouen.fr/cismef/`)

[4]`http://www.inserm.fr/`

3 Methods

4 of the 7 terminologies studied already contain most CUI and TUI information, but poorly encoded as a `skos:altLabel` among the numerous other labels of the classes. For the remaining ontologies, the information had to be found independently either through existing multilingual translation mappings or directly through querying UMLS Metathesaurus through its SQL interface. Our goal is to formally represent CUIs with the `umls:cui` property and TUIs with the `umls:tui` relation, where the `umls` namespace is defined as: `http://BioPortal.bioontology.org/ontologies/umls/`. By using this namespace, the NCBO and SIFR BioPortal can automatically recognize UMLS identifiers and use them properly within the platform services, especially when filtering annotations created by the Annotators. We applied the following algorithm for each class of the ontology (each subclass of `owl:Class`):

1. Query the existing ontology, retrieve all alternative labels and attempt to match a CUI of the form CXXXXXXX with a regular expression, where each X is a digit.

2. If no CUIs were defined as class labels, use multiligual mappings (Annane et al., 2016). If a mapping is found, query the corresponding English language version of the resource in the NCBO BioPortal and retrieve the CUIs.

3. If no mapping is found (or no CUI information), extract code (unique code in the source ontology) either directly through the `skos:notation` relation, when it is available or from parsing the URIs of the classes. Query UMLS through the UMLS SQL interface to retrieve the CUIs.

4. Otherwise, the class remains without CUIs.

Once we obtain all the CUIs for each class (when possible), we retrieve the corresponding semantic types for each CUIs through the UMLS SQL interface and add them to the model through the `umls:tui` property.

We implemented this algorithm in Java, using the Jena library to load the source and target ontologies as well as the mappings. We used the 2015ab version of UMLS loaded on a MySQL server that we accessed through the Java JDBC API. The algorithms were applied on the ontologies one-by-one. The implementation is available on github[5]. Table 1 quantifies the results of the CUI enrichment.

4 Evaluation

An interesting use-case for the enrichment of the French biomedical ontologies from SIFR BioPortal with UMLS CUIs is the evaluation of the named entity recognition performance of SIFR Annotator on the Quaero Annotated Corpus (Névéol et al., 2014).

The Quaero corpus is a French-language corpus in the biomedical domain for the evaluation of named entity recognition and normalization. Quaero is more specifically composed of two sub-corpora, EMEA which contains information on marketed drugs and the MEDLINE corpus,which contains titles from PubMed abstract titles. The annotations consist of token or phrase boundaries of identified entities, the corresponding UMLS semantic groups and one or more UMLS CUIs. A semantic group is a thematic grouping of several semantic types, for example “Disorder” or “Procedures”. The 10 Semantic Groups are often used as coarse-grained groupings of UMLS Semantic Types (McCray et al., 2001).

The corpus was created by instructing bilingual annotators to annotate the French text with UMLS semantics groups and CUIs based on their English language descriptions and definitions as included in UMLS. This process actually biases the corpus, as there is an implicit translation task hidden within the evaluation of the named entity recognition, which creates a disadvantage for a system such as the SIFR BioPortal annotator that annotates directly with French biomedical ontologies rather than using a translation-based approach.

[5] `https://github.com/sifrproject/sifr_project_java_ontology_processing`

Ontology	#Classes	w/o CUI	w/o TUI	In label	In mapng.	Through code	#Remaining w/o CUI	#Remaining w/o TUI
CIF	1496	1496	1495	1495	0	0	1	1
CISP2	745	742	682	682	0	61	2	2
CIM-10	19853	19853	19813	12021	7792	0	40	40
MDRFRE	66382	4	66378	0	0	0	4	4
MSHFRE	27459	4	0	0	0	0	4	4
MTHMSTFRE	1704	4	1700	0	0	0	4	4
MEDLINEPLUS	849	849	795	795	0		54	54
SNMIFRE	106291	106291	102093	96756	5337	127	4071	4071
WHO-ARTFRE	3483	3483	3482	3320	162	0	1	1
ATCFRE	5768	5768	5755	0	0	5755	13	13

Table 1: Statistics for the ontologies enriched in CUIs (all UMLS ontologies with a French-language version): Number of classes, number of classes without CUIs at the beginning, the number of classes without CUIs at the beginning, the number of CUIs found in labels, the number of CUIs found through mappings, the number of CUIs found through UMLS codes, the number of classes remaining without CUIs at the end and the number of ontologies remaining without semantic types at the end.

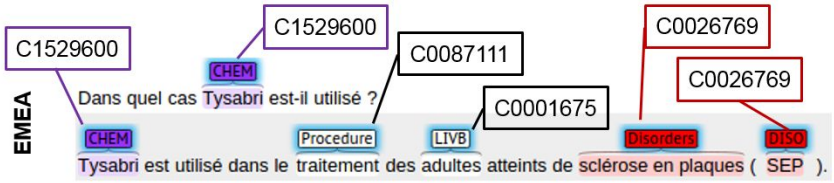

Figure 1: An illustration of the type of annotations in the Quaero corpus. From Névéol et al. 2014.

The evaluation of the named entity recognition is bound to the proper recognition of its semantic group: if the token boundaries (NER) or the CUI identified are correct but the semantic group is incorrect, the annotation is counted as incorrect. This is a confounding factor in the evaluation of NER alone, as the absence of semantic types in a particular ontology will lead to false negatives, although the entity was identified. Figure 1 illustrates the annotations expected in the Quaero corpus.

The SIFR Annotator proposes a specific output format for the Quaero evaluation and several variants. The `quaero` output is the direct output of the annotations as they are returned. The `quaerosg` format is the same, except that when there are several possible semantic groups, the first is chosen. The `quaeroimg` output excludes annotations with ambiguous semantic groups altogether. Although the interface does not show it, the formats can be used through the `format=quaero/quaeroimg/quaerosg` option of the REST API.

Corpus or System	NER + Semantic Groups			NER + Semantic Groups + CUIs		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
EMEA before	7.44	16.47	10.25	6.21	15.36	8.89
EMEA after	69.98	48.61	57.37	42.55	29.32	34.61
MEDLINE before	29.97	57.91	39.50	12.10	24.31	16.16
MEDLINE after	70.06	51.94	59.65	40.87	30.64	35.02

Table 2: The results on the Quaero corpus before and after the CUI enrichment.

We run the evaluation of the SIFR Annotator on the test sets of the EMEA and MEDLINE sub-corpora in Quaero with all the possible UMLS ontologies in SIFR BioPortal. Table 2 presents the compared results. Before the enrichment only MDRFRE, MSHFRE and MTHMSTFRE had CUI information, the lack of CUIs and semantic types prevented the proper annotation and led to very low precision and recall. The fact that the MEDLINE corpus has somewhat better results is due to its good coverage by the MSHFRE, MDREFRE and MTHMSTFRE ontologies. The CUI/TUI enrichment process allowed us to eliminate the precision/recall issue, however errors remain because of ambiguous annotations (a phrase or text generates several annotations where the corpus expects only one). We are now working on addressing these more specific issues with a word sense disambiguation component in SIFR Annotator.

The enrichment in semantic types and CUIs also enables to filtering of annotation results by semantic group with all of the French UMLS source ontologies (Figure 2).

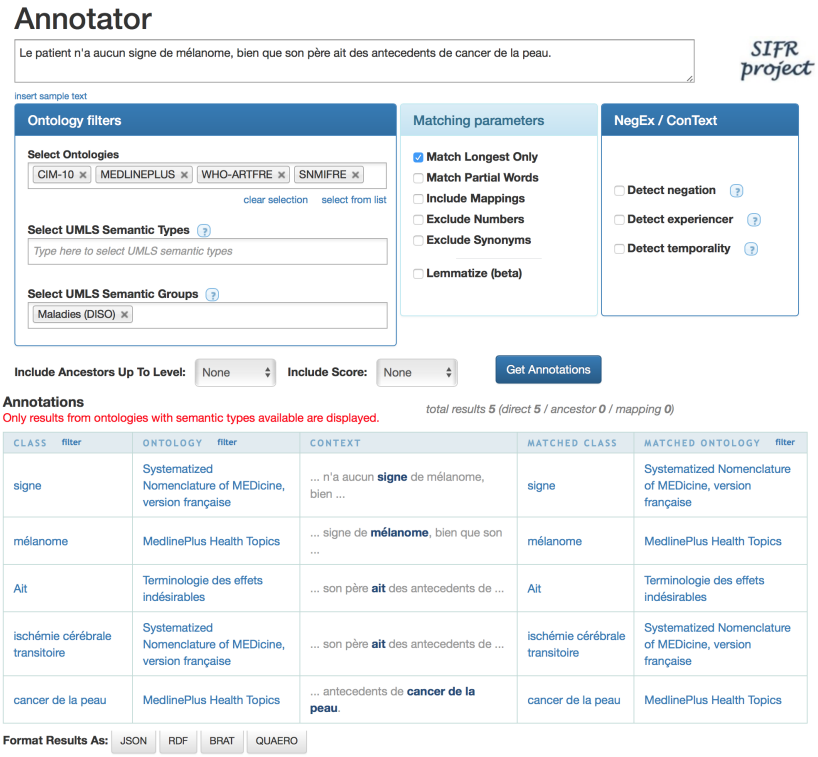

Figure 2: An example of annotation filtering with UMLS semantic types and groups in SIFR BioPortal Annotator.

5 Conclusions and Future Work

We have proposed an approach to enrich French biomedical ontologies in SIFR BioPortal with UMLS CUIs and semantic types in order to improve the annotation performance of SIFR annotator for UMLS based NER tasks. While we achieve our goal on the context of the evaluation on the Quaero corpus, the approach relied only existing mappings and a code interoperability between UMLS and its source ontologies, which is a good start, but does not allow to enrich arbitrary ontologies. The integration of multilingual ontology mapping algorithms into the process may make the small tool we developed for the alignment worthy of integration directly into SIFR BioPortal to allow on-the-fly enrichment whenever a user submits an ontology.

We have described a method to enrich French medical terminologies in the SIFR BioPortal with UMLS concepts and semantic type identifiers in order to improve the annotation performance of SIFR Annotator for UMLS based named entity recognition tasks. While we achieve our goal in the context of the evaluation on the Quaero corpus, the task was relatively easy, but fastidious, as we could rely on

existing multilingual translation mappings and/or a code reconciliation between UMLS sources and the French translated terminologies. Our future perspective is to automatically enable such an enrichment (at least with TUIs) for any ontology uploaded to the SIFR BioPortal. We believe we could rely on knowledge-based ontology alignment techniques to achieve this result.

References

Annane, A., V. Emonet, F. Azouaou, and C. Jonquet (2016). Multilingual mapping reconciliation between english-french biomedical ontologies. In *6th International Conference on Web Intelligence, Mining and Semantics, WIMS'16*, Number 13, pp. 12. ACM.

Bodenreider, O. (2004, 01). The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research 32*(Database issue), D267–D270.

Bodenreider, O., S. J. Nelson, W. T. Hole, and H. F. Chang (1998). Beyond synonymy: exploiting the umls semantics in mapping vocabularies. pp. 815–819. American Medical Informatics Association.

Bollegala, D., G. Kontonatsios, and S. Ananiadou (2015, 06). A cross-lingual similarity measure for detecting biomedical term translations. *PLOS ONE 10*(6), 1–28.

Darmoni, S., E. Jarrousse, P. Zweigenbaum, P. Le Beux, F. Namer, R. Baud, M. Joubert, H. Vallée, R. Côté, A. Buemi, D. Bourigault, G. Recourcé, S. Jeanneau, and J. Rodrigues (2003). Vumef: Extending the french involvement in the umls metathesaurus. *AMIA Annual Symposium Proceedings 2003*, 824–824.

Grosjean, J., T. Merabti, N. Griffon, B. Dahamna, and S. Darmoni (2011, 8–10 November). Multiterminology cross-lingual model to create the european health terminology/ontology portal. In *Short papers of the 9th International Conference on Terminology and Artificial Intelligence, TIA 2011*, Paris, pp. 118–121.

Jonquet, C., A. Annane, K. Bouarech, V. Emonet, and S. Melzi (2016, July). SIFR BioPortal : Un portail ouvert et gnrique dontologies et de terminologies biomdicales franaises au service de lannotation smantique. In *16th Journes Francophones d'Informatique Mdicale, JFIM'16*, Genve, Suisse, pp. 16.

Jonquet, C., V. Emonet, and M. A. Musen (2015, June). Roadmap for a multilingual BioPortal. In *MSW4'15: 4th Workshop on the Multilingual Semantic Web*, Volume 1532 of *CEUR Workshop Proceedings*, Portoroz, Slovenia.

Jonquet, C., N. H. Shah, and M. A. Musen (2009, March). The Open Biomedical Annotator. In *American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09*, San Francisco, CA, USA, pp. 56–60.

McCray, A. T., A. Burgun, and O. Bodenreider (2001). Aggregating UMLS semantic types for reducing conceptual complexity. *Studies in health technology and informatics 84*(0 1), 216.

Merabti, T., L. F. Soualmia, J. Grosjean, M. Joubert, and S. J. Darmoni (2012). Aligning biomedical terminologies in french: Towards semantic interoperability in medical applications. In S. Mordechai and R. Sahu (Eds.), *Medical Informatics*, Engineering Technology in Medicine. InTech.

Névéol, A., C. Grouin, J. Leixa, S. Rosset, and P. Zweigenbaum (2014). The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, pp. 24–30.

Noy, N. F., N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. B. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, and M. A. Musen (2009, May). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research 37*((web server)), 170–173.

Rajput, A. M. and H. Gurulingappa (2013). Semi-automatic approach for ontology enrichment using umls. *Procedia Computer Science 23*, 78 – 83. 4th International Conference on Computational Systems-Biology and Bioinformatics, CSBio2013.

Sarkar, I. N., M. N. Cantor, R. Gelman, F. Hartel, and Y. A. Lussier (2003). Linking biomedical language information and knowledge resources: Go and umls. *Pac Symp Biocomput*, 439–450.

Whetzel, P. L. and N. Team (2013, April). NCBO Technology: Powering semantically aware applications. *Biomedical Semantics 4S1*(S8), 49.

Zweigenbaum, P., R. Baud, A. Burgun, F. Namer, E. Jarrousse, N. Grabar, P. Ruch, F. Le Duff, B. Thirion, and S. Darmoni (2003). Towards a unified medical lexicon for french. *Stud Health Technol Inform 95*, 415–420.