Natural Language Engineering (2023), **29**, pp. 1364–1401
doi:10.1017/S1351324922000055

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

MHeTRep: A multilingual semantically tagged health terms repository

Jorge Vivaldi[1]* and Horacio Rodríguez[2]

[1]Universitat Pompeu Fabra, Barcelona, Spain and [2]Universidad Politécnica de Catalunya, Barcelona, Spain
*Corresponding author. E-mail: jorge.vivaldi@upf.edu

(Received 4 December 2019; revised 16 January 2022; accepted 17 January 2022; first published online 25 February 2022)

Abstract

This paper presents MHeTRep, a multilingual medical terminology and the methodology followed for its compilation. The multilingual terminology is organised into one vocabulary for each language. All the terms in the collection are semantically tagged with a tagset corresponding to the top categories of Snomed-CT ontology. When possible, the individual terms are linked to their equivalent in the other languages. Even though many *NLP* resources and tools claim to be domain independent, their application to specific tasks can be restricted to specific domains, otherwise their performance degrades notably. As the accuracy of *NLP* resources drops heavily when applied in environments different from which they were built, a tuning to the new environment is needed. Usually, having a domain terminology facilitates and accelerates the adaptation of general domain NLP applications to a new domain. This is particularly important in medicine, a domain living moments of great expansion. The proposed method takes Snomed-CT as starting point. From this point and using 13 multilingual resources, covering the most relevant medical concepts such as drugs, anatomy, clinical findings and procedures, we built a large resource covering seven languages totalling more than two million semantically tagged terms. The resulting collection has been intensively evaluated in several ways for the involved languages and domain categories. Our hypothesis is that MHeTRep can be used advantageously over the original resources for a number of NLP use cases and likely extended to other languages.

Keywords: Term extraction and automatic indexing; Natural language processing for biomedical texts biomedical texts; Multilinguality; Resources for biomedical NLP

1. Introduction

For many domain-restricted NLP tasks and applications, having semantically organised lexical resources (terminologies, lexicons, ontologies, etc.) is mandatory. The medical (and more generally, the health) domain includes all activities related to the diagnosis, treatment and prevention of disease, illness, injury and other physical and mental impairments in humans. Health care is delivered by practitioners in medicine, chiropractic, dentistry, nursing, pharmacy, allied health and other care providers. The health care industry is a sector that provides goods and services to treat patients with curative, preventive, rehabilitative or palliative care has a great importance and probably it is the most important domain. Consequently, in the context of this paper, we consider that all documents related, either directly or indirectly, with in this field of activity constitute the medical domain. This paper aims to create a large collection of terms used in this domain in all types of publications and text genres.

This domain is also interesting because of its complicated and difficult linguistic characteristics making the job specially challenging[a]. Athenikos and Han (2010) and Cohen *et al.* (2020) survey these characteristics including: (i) the variety of genres as electronic health reports (*EHR*), discharge and radiological summaries, medical papers, drug leaflets, medical fora (for expert and lay users), clinical trials, etc., (ii) complex terminology, including extremely ambiguous acronyms and (iii) the gap between lay and expert users may be larger than in other restricted domains. For instance, the 'Creutzfeldt-Jakob disease', also named 'subacute spongiform encephalopathy' or 'vCJD', is known at lay level as the 'mad cow disease'. But what is more important: the task of terminology extraction is feasible, though not easy, in the medical domain due to the availability of lexical resources at least for some languages and at least for some types of information.

The aim of our work is obtaining a huge collection of medical terms clustered into coherent semantic classes for a set of languages. Although our hypothesis is that our method could be applied to any language having appropriate resources, we have limited ourselves to seven languages including the most resourced one: English, three medium size: French, German and Spanish and three less resourced: Arabic, Basque and Catalan. Besides the need to include English and Spanish, the reasons for this selection are twofold:

- The first reason is somehow opportunistic: For some languages, we are involved in projects which having such resources available. This is the case of Arabic, Basque and Catalan.
- The second reason is more related to our evaluation framework and to check the adequacy of our approach applied to other languages. We want to cover a wide spectrum of languages according to the existence of resources for each one and to the size of such resources. At one extreme of the spectrum, we can place English for which general resources such as WordNet[b] (*WN*), Wikipedia[c] (*WP*), DBpedia,[d] (*DPP*) and Freebase[e] and medical domain specific, as UMLS[f], MesH[g], Snomed-CT[h], BioPortal[i], etc. exist and are freely available. At the other extreme, we include less resourced languages such as Arabic, Basque and Catalan. In the middle, we include languages owning incomplete resources, of small size or covering only some of the semantic domains. Besides Spanish, we have included French and German.

Please note also that the chosen languages belong to different families and therefore each one has its own linguistics characteristics that have to be taken into consideration when using NLP techniques in such languages. In addition, there is a growing interest in developing new resources for languages other than English and also incorporating terminologies and ontologies in relevant resources such as UMLS and Snomed-CT (Névéol *et al.* 2018). Our choice includes well-resourced languages, as English and, at distance, Spanish and French, with ill-resourced languages, including Arabic for which no specific resource is available.

We need to set clearly what *MHeTRep* is and what it is not. What we have built is a multilingual terminology for seven languages, semantically tagged with a well-defined tagset. Links from the terms included into our terminology to their original sources are recorded. So, in the case of terms collected from more than one source, these links allow indirect links between such sources. The case of multilingual links is especially useful. The origin (the sources) of these terminologies are usually rich structures (databases such as DrugBank or ontologies such as FMA, Snomed-CT,

[a] See the difficulties described in Goodwin *et al.* (2020) for obtaining reference summaries in this domain.
[b] English WN: https://wordnet.princeton.edu/
[c] English WP: https://www.wikipedia.org/
[d] https://wiki.dbpedia.org/
[e] https://developers.google.com/freebase/
[f] https://www.nlm.nih.gov/research/umls/
[g] https://www.ncbi.nlm.nih.gov/mesh
[h] http://www.snomed.org/
[i] https://bioportal.bioontology.org/

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

MeSH and others). So, our proposal is rich from the point of view of coverage but poor from the point of view of the information attached to each term (properties of nodes of an ontology) and the structural information derived from the topology of the ontologies. In some way, the sources and the resulting terminology are complementary.

Using *MHeTRep* offers some advantages over using one or several of the original resources integrated into it:

- Obviously the coverage of whole *MHeTRep* is higher than the coverage of a subset of the integrated resources.
- *MHeTRep* offers a uniform, coherent and easy way of accessing the data. This is not the case of the original resources which have specific ways of access. An exception are the ontologies integrated into *BioPortal* that can be accessed through a common API[j].
- *MHeTRep* terms are semantically tagged using a tagset based on top categories of *Snomed-CT*. The tagsets used by other resources, if any, are different and the mappings between them are not trivial at all.
- *MHeTRep* allows the use of APPROX matching. Although it is possible to implement a similar system for other resources easily, the process has to be repeated for each new resource. Using a unique suffix trees (*ST*) representation for the whole collection is better in terms of saving space, efficient access and cost of development.
- *MHeTRep* includes a relatively rich coverage of tagged terms for less resourced languages. In fact, the multilingual coverage is high.
- *MHeTRep* offers useful mappings between terms coming from different sources, and, so, can take advantage of the information contained in such sources. Of particular interest is the mapping between terms in different languages.
- *MHeTRep* allows an extremely easy way of implementing use cases as the following:
  - looking for the exact/approximate existence of a term of a language overall or in the vocabulary of a specific class. *cat* or in any other vocabulary.
  - looking for translations maps.
  - getting the whole vocabulary for a given language and/or class of a language *l* and a class *cat*.
  - performing a semantic tagging of a medical document, such as an EHR, in a language *l*. An EHR usually contains mentions of diseases, drugs, procedures, body parts, etc. It is possible to look at DrugBank for identifying drugs, at FMA for body parts, at ICD-10 for diseases, etc., but it seems more efficient to perform a unique look up over an integrated resource.

After this introduction, the organisation of the paper is as follows. In Section 2, a brief survey of involved or related disciplines is presented. In Section 3, the semantic tagset and the lexical resources used in our work are shown. The methodology to be applied is proposed in Section 4. This methodology is applied, and the results obtained are shown in Section 5 and evaluated in Section 6. Finally, in Section 7, we present our conclusions and future work. In Table 1, we present the notations most frequently used along the paper[k].

[j] http://data.bioontology.org/documentation

[k] For language-dependent terms, we can use either the generic term with the super-index *l* or the specific one with the corresponding super-index: *ar*, . . ., *fr*, for instance, a generic *WP* for a language *l* can be referred as WP^l, while English *WP* is referred as WP^{en}. Similarly for other super-indexes.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 1. Notation used in this paper

Term	Definition
WP^l	Wikipedia for language *l*
WN^l	WordNet for language *l*
DBP^l	DBP for language *l*
t	Term
tc	Term candidate
tag or *cat*	Semantic tag
tagS or *catS*	Semantic tagset
langS	Set of languages considered
tS^l	Set of terms for language *l*
$tS^{l,s}$	Set of terms for language *l* and source *s*
tS^l_c	Set of terms for language *l* and semantic tag *c*
$tS^{l,s}_c$	Set of terms for language *l*, semantic tag *c* and source *s*
$tS^l_{c.inic}$	Initial set of classified term candidates for category c and language l
$tS^l_{c.extended.i}$	Set of classified term candidates for category c and language l after iteration i

2. State of the art

2.1 Term extraction

Obtaining the terminology of a domain is needed for tasks such as building/enriching lexical repositories, lexicon updating, summarisation, named entity recognition or information retrieval among others. At the same time, it is a problematic task. Two problems arise: first, a well-organised corpus of texts representative of the domain is needed and, second, how the terms of this corpus could be obtained.

Compiling a corpus is an expensive task in time and resources. Obtaining the terms included in such a corpus is also problematic. Manual processing is unrealistic, and thus automatic methods are used, although the results are not perfect.

In the case of lacking of terminological resources in the domain of interest, the use of a term extractor system could help. Usually, terms are defined as lexical units that designate concepts in a thematically restricted domain. As mentioned before, terms are useful for a number of *NLP* tasks. For many of them, term recognition constitutes a serious bottleneck. Since the nineties, this task has been object of research but, in spite of these efforts, it cannot be considered solved. Term extraction can be seen as a semantic tagging task as it adds meaning information to the text.

The way to tackle this task depends on the available resources, mainly ontologies and lists of terms. If these resources are not available, it is necessary to resort to indirect information of a linguistic and/or statistical nature. The results obtained in these ways are limited and therefore these tools tend to favour coverage over accuracy. The consequence is that many extractors get long lists of candidates to be verified manually. One of the reasons for this behaviour is the lack of semantic information.

Due to the lack of semantically tagged resources and as shown in Cabré *et al.* (2001) and Pazienza *et al.* (2005), many indirect methods have been proposed to obtain the terms included in texts. Some of them are based on linguistic knowledge, as in Heid *et al.* (1996), others use

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

statistical measures, such as ANA (Enguehard and Pantera 1995). Some approaches such as TermoStat (Drouin 2003) or (Frantzi *et al.* 1998) combine both. Linguistic methods are based on the analysis of morphosyntactic patterns, but this leads to a noisy result; furthermore, the resulting candidates cannot be scored, leaving the measure of relevance to the evaluation of experts in the domain. Statistical methods, instead, focus on detecting terms using statistical measures, often the simple frequency. Terms candidates may be ranked but again an expert is necessary to measure its actual relevance.

Recently, Lossio-Ventura *et al.* (2016) proposed a huge set of measures based on linguistic, statistical, graphical and web information to evaluate the termhood of set of term candidates. Some of them are new while other are modification of already known measures.

Machine learning methods have also been applied to terms extraction including, by design, both term extraction and term classification tasks. Usually, these methods require huge amount of training data. The lack of reliable tagged resources for training constitutes one of the main issues. Another issue concerns the detection of term boundaries which are difficult to learn. Some examples of these techniques are shown at Conrado *et al.* (2013) and Newman *et al.* (2012). Also, recent techniques like those using deep learning have been applied to the task, as shown in Wang and Liu (2016) and Bay *et al.* (2021).

A common limitation of most extractors is the lack of semantic knowledge. Notable exceptions for the medical domain are MetaMap (Aronson and Lang 2010) and YATE (Cabré *et al.* 2001). Most approaches focus on technical domains in which specific resources are available[l] and term extraction is easier. As documents use to be terminologically dense, term detection is easier. A drawback of many documents in this domain (health records, clinical trial descriptions, events reports, etc.) is that often they include spelling errors, domain/institution specific abbreviations and can be syntactically ill formed.

Recently, deep learning models have achieved great success in fields such as computer vision and pattern recognition among others. *NLP* research has also followed this trend as shown in Young *et al.* (2017). This success is based on deep hierarchical features construction and capturing long-range dependencies in data. These techniques have also been applied in EHR for clinical informatics tasks (see Shickel *et al.* (2018) for a good review).

When several resources are available for extracting terminologies for a domain, some approaches take profit of redundancy for improving the accuracy of the individual extractors. Dinh and Tamine (2011) are an example of such approaches in the biomedical domain using approximate matching for increasing its coverage.

2.2 Combining terminologies in medicine

Medical terminology is especially challenging because of its richness, high level of polysemy and lexical variants. Therefore, applications based on medical terminologies could improve their performance by taking advantage of combining different resources (Smith and Scheuermann 2011). Existing approaches to the combination of terminological resources in the medical domain include both repositories of resources, as *UMLS* or *BioPortal*, allowing the interconnection of independent resources and software facilities for a uniform access, and tool suites, as *PyMedTermino*, that provide software support, as wrappers, for the integration of external resources. Some examples are as follows.

UMLS[m] (Bodenreider 2004). Unified medical language system comprises datasets and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. It includes three main sources of knowledge, although the

[l]Note that most of the available resources refers to English.
[m]https://www.nlm.nih.gov/research/umls/

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

only relevant one in this context is the Metathesaurus. UMLS is probably the richest repository of medical terminology, although it has a serious limitation: it basically is a monolingual English resource. It includes vocabularies, hierarchies, definitions, relationships and attributes but only a fraction of them involve languages other than English. More specifically, UMLS includes 216 resources but only 35% involves languages other than English. Some efforts have been carried out recently for increasing its multilinguality (Hellrich *et al.* 2015), with little success for the moment.

BioPortal[n] (Whetzel *et al.* 2011; Salvadores *et al.* 2012) is a web-based service to give access to commonly used biomedical ontologies, and 873 are currently available[o]. These resources are true ontologies, richer than vocabularies or terminologies. *BioPortal* offers several ways of uniformly accessing its content through a REST API[p]. The format for accessing the content is simple and powerful. A popular service is the NCBO annotator[q] that allows selection of some ontologies for performing semantic tagging of a document. Unfortunately, the resource is basically English monolingual although a few of the included ontologies allow the so-called multilingual views. Some effort has been made to improve multilingualism in the *BioPortal*, for example, the roadmap proposed by Jonquet *et al.* (2015), but currently these proposals are still only a project.

PyMedTermino (Medical Terminologies for Python)[r] (Lamy *et al.* 2006) is a Python module for easy access to an open collection of medical terminologies. Although built for the French language and including only wrappers for French resources, it seems not difficult to extend its coverage to other languages and resources.

3. Lexical resources

In this section, we present the semantic tagset chosen for classifying our terms and the lexical sources used for building the terminologies.

3.1 Defining the tagset

Selecting a tagset is a task that tries to balance the usefulness of the categories and the difficulty of the task of classification. The granularity of the tagset is the most important choice to make. For instance, should we use a unique category for clinical findings or do we have to distinguish among diseases, disorders, symptoms, etc.?

Several choices arise for defining the tagset: defining it from scratch or using categories of already existing medical ontologies, such as UMLS, ICD10 and MeSH. We decided to use an already defined tagset instead of creating a new one from scratch.

The choice of which tagset to use, like any choice, is never without controversy. In this case, the choice was *Snomed-CT* due to its good coverage throughout the medical field, its adequate granularity, its ease of mapping to other resources and its wide dissemination in hospital settings. For this last point, it is worth mentioning that it is used in a number of relevant medical institutions around the world (Mount Sinai and Mayo Clinic in the USA, Hospital Clínico in Europe and HIBA in South America among others).

Table 2 shows the full tagset used in this task together with a short description of each category, taken from *Snomed-CT* documentation[s].

[n] https://bioportal.bioontology.org
[o] Visited 24 May 2021.
[p] http://data.bioontology.org/documentation
[q] https://bioportal.bioontology.org/annotator
[r] https://pypi.org/project/PyMedTermino/
[s] http://www.snomed.org/

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 2. Snomed-CT semantic categories

Snomed-CT Category	Abbrev.	Description
Body structure	BOS	Normal and abnormal anatomical structures
Clinical finding	CLF	Normal and abnormal clinical states
Environment or geographical location	ENV	Types of environments as well as named locations such as countries, states and regions.
Event	EVT	Occurrences excluding procedures and interventions
Observable entity	OEN	Question or assessment which can produce an answer or result
Organism	ORG	Organisms of significance in human and animal medicine
Pharmaceutical/biological product	PHA	Drug products
Physical force	PHF	Physical forces that can play a role as mechanisms of injury
Physical object	PHO	Natural and man-made physical objects
Procedure	PRO	Medical and surgical procedures performed for diagnostic or therapeutic procedures, administrative procedures, etc.
Qualifier value	QUV	Values for some SNOMED-CT attributes, where those values are not subtypes of other top-level concepts
Record artefact	REC	Content created for the purpose of providing other people with information about record events or states of affairs
Situation with explicit context	SIT	Concepts in which the clinical context is specified as part of the definition of the concept itself
SNOMED-CT model component	MOD	Contains technical metadata supporting the SNOMED-CT release
Social context	SCT	Social conditions and circumstances significant to health care
Special concept	SCO	Concepts that do not play a part in the formal logic of the concept model of the terminology, but which may be useful for specific use cases
Specimen	SPC	Entities obtained (usually from the patient) for examination or analysis
Staging and scales	SCA	Assessment scales and tumour staging systems
Substance	SUB	General substances, chemical constituents of pharmaceutical/biological products, body, dietary and diagnostic substances

3.2 Lexical datasets

The lexical resources we have used for our task can be classified along several axes:

- Regarding the organisation of the resource, we include ontologies, thesauri, lexical databases, lexicons, vocabularies, terminologies and annotated corpora.
- Regarding the domain scope, we can consider general, medical and sub-medical specific resources[t]. WN, WP and DBP are examples of the general category, Snomed-CT, UMLS, MeSH and Galen belong to the medical category and FMA, DrugBank, ICD10, CIE and CIM belong to the latter.

[t]As we are interested only on the medical domain, we refer to those resources covering any specific domain (including medicine) as general, those covering the whole domain of medicine as medical and those covering specific aspects of medicine (such as drugs, diseases and anatomy) as domain specific.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

- Regarding the languages covered, we distinguish between monolingual and cross-lingual resources[u]. Examples of monolingual resources are MeSH, Galen, FMA and DrugBank[v] for English. WN, DBP, Snomed-CT[w] and ICD10[x] are cross-lingual.

3.3 Datasets used in developing MHeTRep

The resources we have are listed in Table 3. Most of them are monolingual; therefore, the column 'Language' contains a single language. In the case of multilingual resources (WordNet, Orphanet, DBP) the different languages and sizes are presented in separated rows. In some cases (DrugBank, RadLex) most of the terms correspond to a main language, but there are also small amounts of terms in other languages. In these cases, we include in the 'language' column all the languages, with the main one in bold, and we present in the 'size' column the summation of terms in all the languages. We give in what follows some details on the resources.

SNOMED Clinical Terms. Snomed-CT (Donnelly 2006) is a systematically organised collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. It is considered to be the most comprehensive, multilingual clinical health care terminology worldwide. *Snomed-CT* provides the core general terminology for electronic health records. It includes terms for all the classes in our tagset[y]. *Snomed-CT*, with more than 350,000 concepts, 950,000 English descriptions (concept names) and 1,300,000 relationships, is the largest single vocabulary ever integrated into UMLS. As mentioned in Section 3.1, we have used the top classes of *Snomed-CT* as the tagset for classifying our terms.

DrugBank. This resource (Wishart *et al.* 2006) is a database containing information on drugs and drug targets. *DrugBank* combines detailed drug (i.e., chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e., sequence, structure and pathway) information. We have used version 5.0.11 (Wishart *et al.* 2017) containing 11,002 drug entries. Additionally, 4910 non-redundant proteins sequences are linked to these drug entries. From the database fields, we have used those containing lexical information, namely the **DrugBank ID**, the **Name**, that is, standard name of a drug as provided by its manufacturer, **Brand names**, commercial names used by drug manufacturers, **Synonyms** and **ATC Code**.

CIMA. Clasificación Internacional de Medicamentos[z] is a resource on drug information in Spanish. *CIMA* supports coding of several tags describing its entries. The tagsets include *ATC* codes, forms of presentation, forms of administration, laboratories, containers, excipients, doses, quantities, units, etc.

The denomination of *CIMA* entries is extremely low level and, so, some higher level denominations of the products have to be derived from the original entry. For instance, a *CIMA* entry is 'SERACTIL 400 mg COMPRIMIDOS RECUBIERTOS CON PELICULA'[aa].

From this entry, we obtain as well 'SERACTIL 400 mg COMPRIMIDOS', 'SERACTIL 400 mg', 'SERACTIL 400' and 'SERACTIL'. In this denotation, '400' has been recognised as a <QUANTITY>, "mg" as a <UNIT>, "COMPRIMIDOS" as a <PRESENTATION> and "RECUBIERTOS CON PELICULA" as a <FORM>. Using these semantic tags as terminals, we have manually built a regular grammar for analysing the decomposition of the denomination. The whole grammar consisted of 67 rules. Obviously, the new terms generated by the grammar

[u] We can also consider the multilingual resources for which different datasets exist for different languages without explicit links between them. None of our resources can be classified into this category.

[v] In fact, DrugBank, although an English resource, includes a small set of drug terminology in French, German and Spanish.

[w] The international version of Snomed-CT exists for several languages. We have used the official English and Spanish releases. For Catalan, we used a translation in progress.

[x] Named 'Clasificación Internacional de enfermedades', CIE-10, for Spanish, https://eciemaps.mscbs.gob.es/ecieMaps/, and 'Classification Internationalle de Maladies', CIM-10, for French, https://www.cihi.ca/fr/.

[y] See the whole list in Table 2.

[z] International Classification of Drugs.

[aa] SERACTIL 400 mg, film-coated tablets.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 3. Resources used in the work

Resource	Description	Langs	Size
FMA	Foundational Model of Anatomics	en	174,329
Snomed-CT	Systematised Nomenclature of Medicine – Clinical Terms	**en**, ca, es	665,940
Galen	Generalised architecture for languages, encyclopedias and nomenclatures in medicine	en	39,264
MeSH	Medical Subject Heading	en	205,522
WP	Wikipedia (Accessed through DBP)	–	–
DBP	DBPedia	en	76,336
		es	15,739
		fr	21,359
		ca	11,448
		eu	2539
Wn	WordNet	en	19,966
		es	14,543
		ca	10,861
		eu	5096
DrugBank	DrugBank database	**en**, de, es, fr	15,867
Orphanet	Lists of rare diseases and orphan drugs	en	2206
		es	1792
		fr	2099
		de	1814
DO	Disease Ontology	en	21,273
ICD-10	International Classification of Diseases, 10th edition	**en**, es, fr	142,000
CIMA	Medicine On-line Information Center of AEMPS	es	32,039
RadLex	Radiology Lexicon	**en**, de	2626

are not error free, that is, there could be false positives. The design of the grammar was, however, very conservative and the ratio of false positives is below 2%.

FMA. Foundational Model of Anatomy (Rosse and Mejino 2003; Mejino *et al.* 2003) is 'an ontology of biomedical informatics concerned with the representation of classes or types and relationships necessary for the symbolic representation of the phenotypic structure of the human body'.

Although created in Protégé, there are ontological representations available in OWL or RDF/XML (Noy *et al.* 2004). These representations are complete but contains information useless

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

for our purpose. *BioPortal* provides a simplified CSV table of FMA's ontological terms. This flattened representation is very suitable for our purposes. To further improve the utility of this tabular FMA representation, Michael Halle[ab] has converted the CSV format into an SQLite database. This is the version we have used in our project.

MesH. Medical Subject Headings (Lipscomb 2006) is a controlled vocabulary for indexing and searching biomedical literature. Its main purpose is to provide a hierarchically organised terminology for the indexing and cataloguing of biomedical information. *MeSH* terms and subheadings are used to indicate the topics of an article. Currently, there are 28,489 descriptors in MeSH 2017, the version we have used.

Galen. Generalised Architecture for Languages, Encyclopedias and Nomenclatures in medicine common reference model (Galen CRM) is a part of the *GALEN* project (Rector *et al.* 2009). It is the pioneer of the use of formal logic in biomedical terminologies (de Freitas *et al.* 2009). It is organised as a set of is-a hierarchies containing overall about 25,000 nodes (concepts) using 26 link types (relations) (Corsar *et al.* 2009). *OpenGALEN* uses GRAIL, a description logic based language (Galen Representation And Integration Language). Implementations in *OWL* and *RDF* are also provided.

ICD-10. The International Classification of Diseases, revision 10, is the most widely used version of ICD. It is translated into 42 languages. Its nodes are structured into 22 chapters (Infections, Neoplasms, Blood Diseases, Endocrine Diseases, etc.) and denote about 14,000 classes of diseases and related problems. It is splitted into *ICD-10-CM* for diseases and *ICD-10-PCS* for procedures. *ICD-10-CM* codes are organised hierarchically, where top-level entries are general groupings and bottom-level codes indicate specific symptoms or diseases and their location. *ICD-10-PCS* contains procedure codes organised as a taxonomy of up to 7 levels with 17 codes at first level.

WordNet. Due to its success, although it has been initially developed for English, wordnets for many other languages have been built. Such borders are defined as those synsets whose (direct/indirect) hyponyms belong (with some confidence degree) to the medical domain while their hypernyms are outside this domain. Therefore, such synsets constitute a border between in domain and out of domain areas in this resource. For example, there is a synset in *MCR* that localises as 'body substance', this implies that all its hyponyms are all different types of substance located in the human body.

In this way, we obtain all the existing nouns in the medical domain. But it is also possible to obtain related words using other relations in MCR. We obtain adjectives following the relation 'pertains_to' and verbs using the relation 'related_to'. The former relation allows to relate the adjective 'bronchial' with 'bronchus' while the latter connects 'injection' with the verb 'to inject'.

RadLex. Radiology reports are probably the most studied type of clinical narrative. They contain terminology not only from the medical domain but also from the domain of imaging and graphical software. The coding system of RadLex is the Index for Radiological Diagnoses (*ACR Index*) that offers both anatomic and pathological identifiers.

Orphanet. Orphanet (rare diseases and orphan drugs) is a reference European resource initiated by the INSERM (French National Institute for Health and Medical Research) and devoted to gather and improve knowledge on rare diseases and orphan drugs. It aims to improve the diagnosis, care, treatment and information for patients with rare diseases.

Although many of the entries of the list of rare diseases are already mapped to Snomed-CT, our interest in this resource is to provide completeness and mapping to different languages and sources.

[ab] https://github.com/mhalle/fma-sqlite, no reference is provided.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 4. Fragment of the entry *angiosarcoma* within DO ontology

tag	Code	Value
id:	DOID	0001816
name:	–	angiosarcoma
alt_id:	DOID	267
alt_id:	DOID	4508
synonym:	–	'hemangiosarcoma'
xref:	MESH	D006394
xref:	NCI	C3088
xref:	NCI	C9275
xref:	SNOMEDCT	33176006
xref:	SNOMEDCT	39000009
xref:	UMLS_CUI	C0018923
xref:	UMLS_CUI	C0854893
is_a:	DOID	175 ! vascular cancer

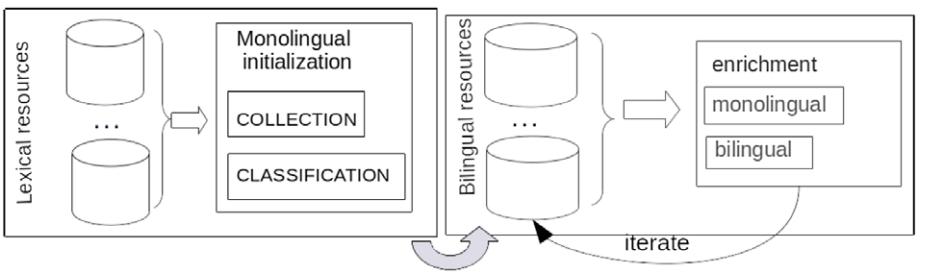

Figure 1. Compilation methodology.

DO. The Disease Ontology (*DO*) has been developed as a standardised ontology for human diseases. *DO* integrates disease and medical vocabularies through extensive cross mapping of *DO* terms to MeSH, ICD, Snomed-CT and other resources. In fact, the reason for including this ontology into *MHeRep* is not the coverage of the diseases, satisfied enough by other resources but the very rich coverage of mappings into other resources. For instance, in Table 4 the entry *angiosarcoma* has one main *DO* identifier, two alternate *DO* identifiers and seven mappings to other ontologies. *DO* contains currently 8043 human diseases.

4. Compilation methodology

Our aim is to obtain semantically tagged medical terminologies for the seven involved languages using the tagset defined in Section 3.1. The structure of our system is depicted in Figure 1. There are two main processes:

- Building the initial monolingual vocabularies.
- Enrichment. Including both monolingual and multilingual enrichment. This second module is applied iteratively until no further enrichment is produced.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

The first component is applied independently for each language and semantic class for which lexical sources exist. For each resource, the process involves two steps: (i) obtaining the set of terms and (ii) classifying them into one or more semantic classes. The first step depends on the organisation of the data source (ontologies, lexical databases, lexicons, etc. represented on OWL, RDF, XML,.csv plain text, databases, etc.).

The classification step is more challenging and will be discussed in Section 4.1. Ideally, the assignment of a term to a semantic class should be unique, but this is not always the case. Not only the assignment criteria for manually curated resources could differ, but even for the same resource the assigned class for a term could be non-unique. Just a couple of examples for illustrating this case: in Snomed-CT, chemical compounds, molecules or atoms are classified as 'Pharmaceutical/biologic product' or 'Substance' depending on their use in Pharmacology. Obviously, this use can change along the time, and also the pharmacological use could be dubious for a human annotator even in the case of good expertness. Body fluids, such as *blood*, *bilis* or *semen*, are in some anatomical resources defined as 'Body structure' while in other resources are classified as 'Substance'. The broad class of 'Clinical findings' included in Snomed-CT the concepts of Disorder, Disease and Symptom while finer distinctions occur in MeSH or Galen. These (and other) disagreements are the origin of the main issues in the classification step of the process for resources as MeSH or Galen, as we discuss in the following subsections.

We organised the acquisition process starting with the most confident resources in terms of both term collection and classification. In that way, when faced with the incorporation of the more challenging ones, we could have some terms already included as learning material. For instance, in DrugBank the extraction process is straightforward because it is organised as a database and the classification is trivial (all the terms are supposed to be 'Pharmaceutical/biologic product'). FMA is also simple because all terms are English and belong to a unique class as 'Body structure'. Once the most confident resources have been processed and their classified terms incorporated to our vocabularies, we can use this material as a support for performing the collection and classification tasks from the more challenging resources.

We proceed, thus, for our first component in the following steps:

(1) We include, first, monolingual category-specific datasets for which no classification process is needed as FMA and DrugBank.

(2) Next, we include multilingual category-specific datasets for all the languages for which the resource exists, such as ICD-10 for English and their Spanish and French counterparts (CIE-10 and CIM-10). In the case of Snomed-CT, the classification process is simple because the categories correspond to the tops of a curated taxonomy.

(3) DO, RadLex and Orphanet are incorporated with no problems.

(4) Next, we process the most challenging datasets, MeSH and Galen, both presenting issues for classification and the former for acquisition too. See details in Section 4.1.

(5) Finally, the different terminologies are integrated into a unique resource. This process implies normalisation, removing (but mapping) duplicates and getting coding when available from the source. All the mapping and coding information is attached to the term. The representation of the integrated lexical structure is presented in Section 5.1.

The results of the first phase are for each language l and each resource s a set of terms $tS^{l,s}$ classified into the appropriate semantic class c (in *tagS*). In this way, we obtain the initial set of classified terms for the language l. Let us name this set as $tS^{l}_{c.inic}$.

Once completed this phase, we enter into a step of cross-lingual enrichment; this process is applied iteratively until no new enrichments are obtained. The process is detailed in Section 4.2. Basically, each iterative step proceeds into two enrichment processes: one monolingual, using as

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

a knowledge source the corresponding DBP and the another, multilingual, using all the available cross-lingual resources:

(1) Enrich $tS^{l}_{c.inic}$ using generic monolingual resources, namely *DBP* and *WN*, for getting extended sets of tagged terms. Let us name this set as $tS^{l}_{c.extended.0}$.
(2) Transfer these extended sets to other languages using multilingual resources. This process is iterative. After each iteration i, for each language l and each class c, a new set of terms is collected. The new sets are named $tS^{l}_{c.extended.i}$ for increasing values of i. In our experiments, no more than six iterations were required to reach a stable state. The cross-lingual resources we have used for obtaining mappings across languages are: (1) *DBP*. Using *label* and *same_as* properties, (2) *WP*. Using interwiki links and (3) *WN*, Using inter-lingual index (*ILI*) links from *MCR*.

4.1 Process of acquisition of original sources

As mentioned above, the first phase of our process consists of the acquisition and classification of vocabularies coming from our original resources.

We briefly described in Section 3.3 the datasets we have used for this step. Table 2 shows a summary of each resource. Then, in Section 4.1, we describe the acquisition and classification processes from these datasets.

The initial datasets, that is, those that can be obtained and classified without any additional support, can be carried out with simple procedures. Next, we face two challenging tasks: the acquisition and classification of MeSH and Galen. These two datasets pose a number of difficulties that are addressed below in this section.

Snomed-CT:. We performed the task for English, Spanish and Catalan. The acquisition task was straightforward because we had the ontologies of the resource and from them locating the lexical nodes was easy. For classifying the lexical items, we navigated the ontology from the lexical nodes through the hypernym edges until reaching the top classes of the ontology. Ideally, the organisation should be a taxonomy but this was not always the case, sometimes a node had more than one parent and, so, the structure was a bit tangled. In these cases, we incorporated the ambiguous lexical items into all the accessible top classes.

DrugBank:. Next, we incorporated DrugBank. The process was also simple. We obtained the lexical items corresponding to names, synonyms, brand names, etc. Although the database is basically for English, some synonyms are tagged with other languages (French, German or Spanish). All the collected items were directly classified as a 'Pharmaceutical/biologic product'.

CIMA:. As mentioned in Section 3.3, the acquisition of the terms from CIMA needs the generation of partial denotations using a regular grammar. The rest of the tasks (extraction and classification) are simple.

FMA:. We accessed the FMA-lite representations using the *sqlite3* python module[ac].

ICD:. The process of incorporation of English ICD-10 and its French and Spanish counterparts is straightforward. ICD-10-CM for diseases[ad] and ICD-10-PCS[ae] for procedures are processed in the same way.

DO:. As mentioned above, we selected *DO* not for enriching our terminologies but for the richness of mappings from each term into other resources. The data are encoded into lines consisting on a tag and a value. As all the terms are diseases, the task is easy.

[ac] https://www.sqlite.org/index.html
[ad] https://www.cdc.gov/nchs/icd/icd10cm.htm
[ae] https://www.icd10data.com/ICD10PCS/Codes

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 5. Top categories of RadLex mapped into our tagset

RadLex top	Snomed-CT category
Anatomical entity	Body structure
Clinical finding	Clinical finding
Non-anatomical substance	Substance
Object	Physical object
Procedure	Procedure
Procedure step	Procedure

Orphanet:. The data are directly available on the web for each language as a simple list of terms although not all the information is directly downloadable. To include these data in our resource was simple because we know in advance that all terms should be classified as 'Clinical finding'.

RadLex:. We have used the RDF version of the resource. From the 15 top categories occurring in it, we reduced our obtention process to the classes shown in Table 5 that are both relevant and easy to map to our tagset. We got an ontology of 850,631 triples. We selected all the nodes of type LEXICAL: 390,556 nodes from which 252,075 are literals. There are 98 relations defined in the ontology but we only use those supporting taxonomic links such as 'subclass' and equivalence links 'equiv' and those related to naming, that is, 'name', 'preferredName', 'preferredNameGerman', 'synonym', 'synonymGerman' and 'umlsTerm'. There are also some useful properties attached to nodes, such as the 'language' property.

Although RadLex is basically an English resource, there is a small set of links to gain access to the translation to other languages. The coverage is (1) English: 109,763, (2) German: 46,560, (3) Latin: 4123, by default assigned to all languages having the Latin script, and (4) none: 91,629, by default assigned to English.

MeSH:. MeSH is a knowledge structure organised from the skeleton of a taxonomy associated with the MeSH Headings. Table 6 presents the top level of the MeSH Headings. Table 7 shows a fragment of this taxonomy. Beyond the MeSH Headings, entries contain a huge number of properties that can be used for classification. The set of allowable qualifiers can be consulted at the MeSH site[af]. Selecting the MeSH entries can be done straightforwardly but mapping the MeSH entries into our tagset is far from easy because of two issues:

If we look at Table 6, it seems intuitive that some top (one character) MeSH headings can be directly mapped into Snomed-CT top categories.

Mapping 'A' ('anatomy') into 'Body structure', 'B' ('organisms') into 'Organism', 'C' ('diseases') into 'Clinical Findings' or 'D' ('drugs and chemicals') into 'Pharmaceutical/biologic product' seem to be quite safe but, unfortunately, this is not always the case. We have computed the distribution of MeSH terms already classified in our vocabularies from the resources applied so far. There are 112 terms having MeSH Heading starting with 'A'. Of these, 50 have been classified as 'Substance', 43 as 'Body structure' and the rest fall into up to 6 other categories. So, classifying all 'A' as 'Body structure' will fail 57% of the cases. Better results are produced in cases 'B' (from 1130 terms, 1071, that is, 95%, have been classified correctly as 'Organism') and 'C' (from 1483 terms, 1414, also 95%, were correctly classified as 'Clinical finding') while in case 'D' from 5584 terms, the majority, 3224, have been erroneously classified as 'Substance' and only 2354 correctly as 'Pharmaceutical/biologic product'. For the other cases, but 'F', the results are also bad. So another approach should be followed.

[af] https://meshb.nlm.nih.gov/

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 6. MesH descriptor categories

Descriptor	Category
A	Anatomy
B	Organisms
C	Diseases
D	Drugs and chemicals
E	Analytical, diagnostic and therapeutic techniques and equipment
F	Psychiatry and psychology
G	Phenomena and processes
H	Disciplines and occupations
I	Anthropology, education, sociology and social phenomena
J	Technology, industry, agriculture
K	Humanities
L	Information science
M	Named groups
N	Health care
V	Publication characteristics
Z	Geographicals

Table 7. MesH Heading examples: fragment of the taxonomy headed by nervous system

MesH Heading	Name
A08.186.211.132.659.237.364	Inferior colliculus
A08.186.211.132.659.237	Corpora quadrigemina
A08.186.211.132.659	Mesencephalon
A08.186.211.132	Brain stem
A08.186.211	Brain
A08.186	Central nervous system
A08	Nervous system
A	Anatomy

We conducted experiments using the second and third levels of the MeSH Headings, and some cases were solved satisfactorily but not all. So, we decided to learn a set of 19 binary classifiers (one for each class). After running the binary classifiers, a voting mechanism was activated for getting the final class if possible. The candidates that could not been assigned were simply rejected. So the number of terms in our datasets coming from MeSH is smaller than the full coverage of this resource. We admit this drop on coverage for assuring an accurate assignment to the tagset. A similar limitation occurs in the case of Galen.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

We used as classifier a simple multilayer perceptron followed by a SoftMax layer. The classifiers were implemented using the Keras environment[ag]. The learning features consist of the extremely rich information contained in MeSH entries, while the supervision is provided by terms present in MeSH and already classified by previous likely accurate sources. The positive examples for each class are those classified in the corresponding class while the negative examples are randomly chosen from those classified in the other classes. The proportion of positive and negative examples is maintained.

Galen:. This source is organised into 83 files (chapters). We have used the RDF implementation of the ontology[ah]. Each file contains a separate ontology but we have integrated all of them into a single one. We have used the 'Lexical' nodes for obtaining the terms and the others for navigating the ontology. From the various relations included, we have used the 'subClassOf' for navigating the taxonomic links, and 'EquivalentClass' for synonymy links. An issue for selection is that the lexical items have a peculiar notation. For instance, 'presencewhichisExistenceOfHyperthyroidism' should be normalised into 'Hyperthyroidism' and mapped into 'Clinical finding'. For performing the normalisation, we manually built a set of about 100 regular transformation rules. The performance of this grammar is not error-free but, by favouring precision over recall (terms occurring only once are rejected), we have limited the ratio of false positives to be under 2%.

For learning the classifier, we built a system similar to the one described for MeSH using all the already classified terms for learning, including MeSH terms in this case. As Galen entries do not have the rich property framework of MeSH, we used as features the original and normalised forms of the entry and its local environment (edges and nodes adjacent to the one to be classified, the hypernymy path from the node to the top of the hierarchy and the file from which the node has been selected).

4.2 Multilingual extensions

We performed monolingual and multilingual extensions using *DBP*, *WN* and *WP*. Below, we briefly describe these processes.

4.2.1 Multilingual extension using DBP

All the languages but Arabic[ai] from our set of languages have a *DBP* repository accessible through a corresponding endpoint using SPARQL queries. So for the bilingual enrichment based on *DBP*, only these languages can be used as sources. We take advantage of the rich ontological structure of the *DBP*. Consider, for instance, the class 'AnatomicalStructure' that can be mapped into our category 'Body structure'. This class owns the following direct hyponyms: 'Artery', 'BloodVessel', 'Bone', 'Brain', 'Embriology', 'Ligament', 'Lymph', 'Muscle', 'Nerve' and 'Vein', nicely covering the main subcategories of 'Body structure' in Snomed-CT ontology.

We have checked the other categories in the same way to assure the appropriateness of the resource. For the multilingual extension, we have used two properties that exist for some of the *DBP* resources, namely 'same_as' mapping a resource in the *DBP* of one language into the corresponding in another one, and '_label', mapping a resource into denominations in other language.

4.2.2 Multilingual extension using WN

All languages involved in MCR are identified by an ILI that allows to potentially obtain the equivalents for all languages. All the equivalents are created during the enrichment process, but its

[ag] https://keras.io/

[ah] http://www.opengalen.org/sources/sources.html

[ai] There is an Arabic endpoint, but it does not seem to work currently.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 8. Size of vocabularies for each language and category after the initial step

Snomed-CT category	ar	ca	de	en	es	eu	fr
BOS	3118	4065	6201	256,076	62,456	2223	6111
CLF	3866	3239	46,023	453,823	266,898	1991	48,590
ENV	0	30	0	3952	3619	14	549
EVT	420	986	641	9149	8535	689	662
OEN	0	200	0	20,045	17,833	46	259
ORG	2223	2423	3057	75,347	69,510	2024	6231
PHA	2353	2667	12,349	157,321	107,938	1129	11,097
PHO	0	401	0	31,390	18,121	258	460
PRO	546	1866	861	289,576	108,630	443	994
QUV	0	373	0	21,051	19,380	103	1,066
SUB	2486	4429	12,663	230,735	61,806	2142	11,386
Other categories	0	22	0	25,071	11,049	3	462
Total	15,012	20,701	81,795	1,573,478	573,912	11,065	44,136

insertion is not mandatory and depends on the ambiguity of the candidate. All the variants of all the synsets of different languages linked to the selected ILI are included into the corresponding terminologies.

4.2.3 Multilingual extension using WP

In the case of terms described as *WP* pages, there are sometimes interwiki links mapping the page into an equivalent one in the *WP* of other languages.

5 Collecting the data

After executing the first step of the process for all the languages and categories, we obtained the results presented in Table 8. Figure 2 summarises it presenting the plots for the different languages and semantic classes. In abscissas, we present the different categories. Ordinates represent the sizes of the vocabularies. Figure 3 puts together the results integrating all the categories. As can be seen, most of the terms belong to the English vocabularies. The reason is obvious as most of the sources are English monolingual. At high distance, we find the Spanish vocabularies (basically coming from Snomed-CT, CIE-10 and CIMA). The representations of the other languages are marginal.

We iterated six times over the second step for enriching the initial terminology. The final results are presented in Table 9. Figure 4 presents, using a logarithmic y scale, the progression of cross-lingual enrichment along the successive iterations.

In Table 10, we present the information obtained for the English term *Liver* within *MHeTRep*, coming from different sources. We split it into two blocks. The first includes some general information: the original source, its Snomed-CT category and the identifiers in all resources where this term is available while the second includes all the available translations into other languages.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

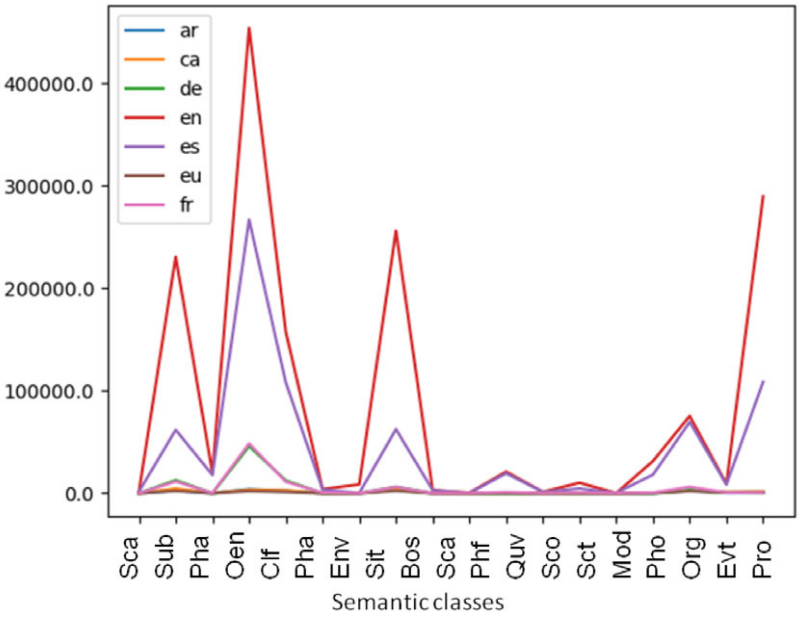

Figure 2. Sizes of first step overall vocabularies per semantic class.

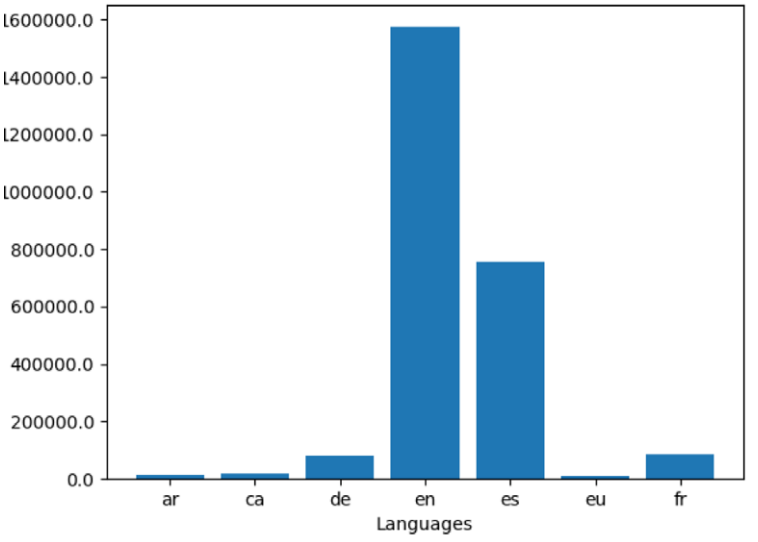

Figure 3. Overall sizes (per language) of the vocabulary after the first step.

5.1 MHeTRep implementation

All the system has been built using *Python*. The whole *MHeTRep* repository is represented as an instance of the class *VOCABULARY_COLLECTION* that can be loaded from and saved to a *Json* file[aj].

[aj] The current final version of the repository as well as a minimal *Python* script for accessing it will be made public.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 9. Size of the resulting vocabulary for each language and category after seven iterations

Snomed-CT category	ar	ca	de	en	es	eu	fr
BOS	3164	4065	6201	256,179	62,498	2245	6175
CLF	3866	3239	46,023	453,827	266,898	1991	4859
ENV	0	30	0	3,952	3,619	14	549
EVT	420	986	641	9,150	8535	689	662
OEN	0	200	0	20,045	17,833	46	259
ORG	2244	2423	3057	75,394	69,539	2041	6266
PHA	4464	2667	13,350	157,324	108,781	1636	12,095
PHO	0	401	0	31,390	18,121	258	460
PRO	547	1866	861	289,576	108,630	443	994
QUV	0	373	0	21,051	19,380	103	1066
SUB	5406	4429	15,215	230,778	63,885	2808	13,885
Other categories	0	22	0	25,013	11,049	3	462
Total	20,111	20,701	85,348	1,573,679	758,768	12,277	47,732

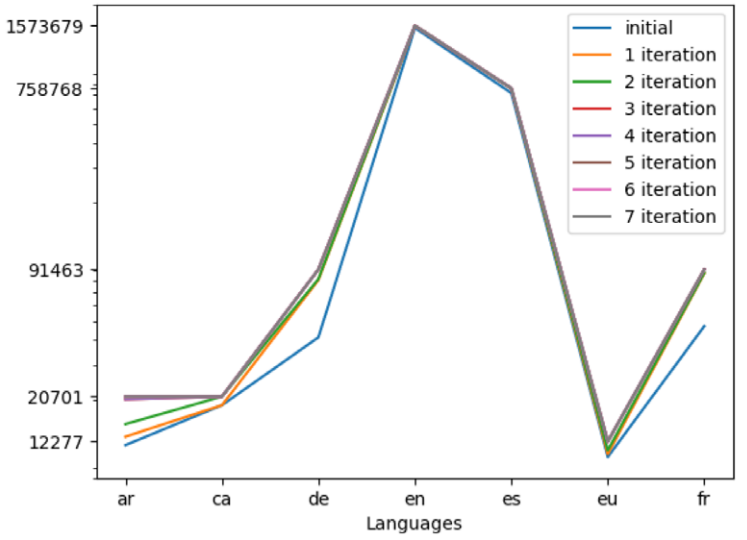

Figure 4. Total size of vocabularies after each iteration step.

From the *VOCABULARY_COLLECTION*, a specific vocabulary for one language and one category can be extracted. Each of these vocabularies is an instance of the class *VOCABULARY* allowing querying for the existence of a term. Each term existing in the *VOCABULARY* has associated an instance of the class *LEXICAL_ENTRY*, containing information of the sources, coding, mappings to other languages, etc.

Managing of these information includes querying, updating and deleting term entries. To allow updating of the data, in case of new versions of the resources, a simple mechanism is also provided that allows to incorporate to *MHeTRep* the tagged terms included in a text file.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 10. Information available for the English term *Liver*

Item	Value
Word	liver
Language	en
Original source	fma
FMA ID	7197
Galen ID	G13751
Snomed-CT ID	10200004
Snomed-CT category	Body structure
dbpedia_labels_en	http://dbpedia.org//resource/Liver
MCR_es ID	spa-30-05385534-n
MCR_en ID	eng-30-05385534-n
MCR_ca ID	cat-30-05385534-n
MCR_eu ID	eus-30-05385534-n
en	liver
es	hígado
fr	foie
de	leber
ca	fetge
eu	gibel
ar	كبد

6. Evaluation

This section presents the evaluation of *MHeTRep*. Evaluating a set of terms, whether they are extracted from a text or, as in this case, obtained through a compilation from different sources raise always a difficult problem. As the process of building *MHeTRep* includes two tasks: getting the terms and classifying them, both tasks have to be evaluated: a term is correct if it is a true term of the medical domain and it has been assigned to the correct category. The evaluation has been done using two different approaches:

(1) Intrinsic evaluation: measuring, for a term candidate t, the confidence of considering t as a term and of classifying it as a class c.

(2) Extrinsic evaluation: using the annotated terms collected from external sources in the domain.

Typically, evaluation of detected terms against reference terms is done by looking for exact matching. It could happen, however, that although some candidates do not match exactly a given reference term, they are not a clear mistakes (e.g. 'neuromuscular blocking agents' or 'chemokine blocking agents' vs. 'blocking agents'). In such cases, using an approximate matching technique can help to take this factor into account. This is why some of the test sources for evaluation includes both EXACT and APPROX matching.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

In what follows, we present the algorithm used to perform the approximate matching, then the details of both types of evaluations and finally, we will draw some conclusions from the results of the evaluation.

As approximate matching is only marginal in our work, we have used a very basic string matching algorithm. Other approaches in the same line are surveyed in Gomaa *et al.* (2013). More recent approaches, as Batet and Sánchez (2020), use ontological context (trees and graphs) of the concepts referred by the terms. Also embedding approaches have been applied with success (Lastra-Díaz *et al.* 2021); for instance, have launched the *HESML* library[ak].

6.1 Approximate matching

A basic component of systems for processing biomedical documents is the recognition and tagging of biomedical entities. For this task, many lexical resources are available and our claim is that the one presented here could be useful. A main problem is that some entities may have a complex word form. For instance, the following entries have been randomly extracted from our terminology coming from the MeSH ontology.

- 2-(hydroxymethyl)-N,N-dimethylpiperidinium benzilate (classified as 'Pharmaceutical/biologic product')
- Interstitial cell of Cajal like cells (a 'Body structure')
- Deficiency disease, betalipoprotein (a 'Clinical finding').

It is extremely unlikely that these forms occur verbatim in biomedical texts. Additionally, (partial) abbreviations and spelling errors are frequent in free text like EHRs, that are the main source of documents for using our terminologies. Also inflectional (as plural forms) and derivational (as adjectival) variants frequently occur. So, some kind of approximate matching is needed.

We have examined the histogram of the length of the terms in the vocabulary of 'Body structure' for English. Figure 5 shows the distribution of length of the terms in characters with a maximum around 40 characters. The distribution of length of the terms in tokens shows a similar shape with a maximum around 5 tokens, so most of the terms are multi-word.

The MHEtrep includes 12,439 atomic tokens[al]. The type/token distribution of the atomic terms is shown in Figure 6, note that the x-axis is logarithmic. The distribution is extremely skewed, close to a Zipfian. 3690 (30%) of the terms occur only once, 9215 (74%) occur less than 10 times and 11,429 (92%) occur less than 100 times. In the other extreme, the terms 'right', 'structure', 'left', 'nerve', 'entire' and 'of' occur more than 15,000 times.

With the character and token histograms in mind, we have built an approximate matching detailed in this section.

For implementing an approximate matching system, a simple exhaustive access to the whole content of the vocabularies is too costly in time. A reasonable alternative could be representing the whole vocabulary (or a fragment) as *ST*, as shown in Figure 7. This compact representation allows an efficient access to both exact and approximate matching.

For this purpose, we used the package 'ST'[am]. Our implementation of approximate matching using *ST* is very simple because we are interested in this task only for the purpose of evaluation. It is worth noting that replicating the algorithm is easy for a specific resource. Our advantage is that we can apply the algorithm over a *ST* representation of the whole collection and the alternative could be using different representations for each resource.

For a given language and a semantic category, we built three vocabularies: the vocabulary of simple terms (mono-word), the vocabulary of complex terms (multi-word) and the vocabulary

[ak] https://github.com/jjlastra/HESML/releases

[al] Individual components of MW terms only, that is, excluding single word terms.

[am] "https://code.google.com/archive/p/suffixtree/"

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

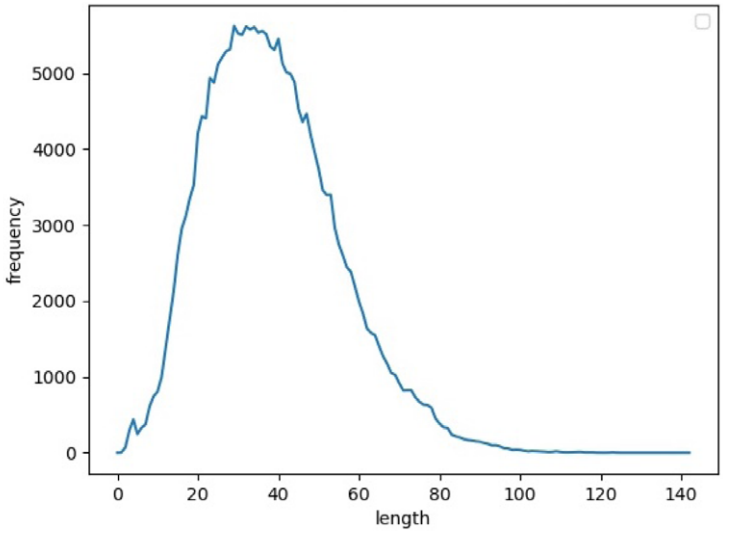

Figure 5. Histogram of atomic terms length in tokens.

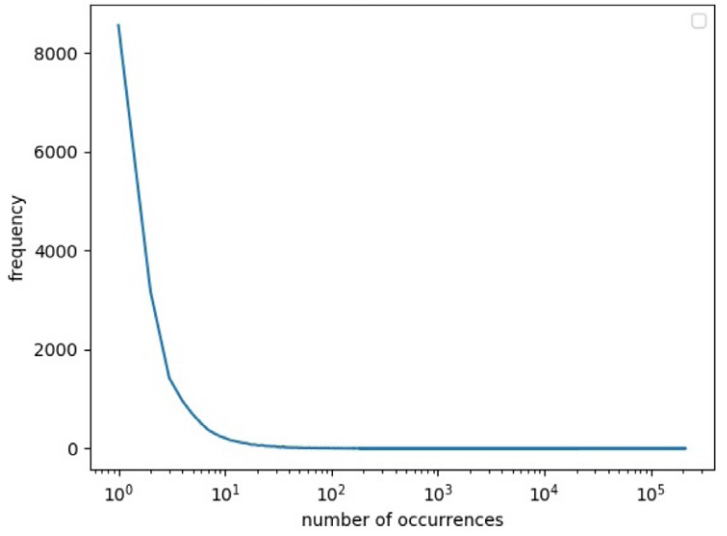

Figure 6. Type/token distribution of the atomic terms.

of atomic terms (atomic components of complex terms). For all these vocabularies, we built a corresponding *ST*. Over this infrastructure, we built a set of approximate extraction rules that we summarise below.

- **Look for approximate extension**: We obtain all the character n-grams (larger than 2) in the term candidate. We look, then, in the simple and complex STs, for all the occurrences of each n-gram. For each occurrence found, we obtain the containing word and we compute the string distance between these words and the candidate. For this purpose, we used the python-string-similarity library[an]. We selected as approximate matched the

[an] https://github.com/luozhouyang/python-string-similarity.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

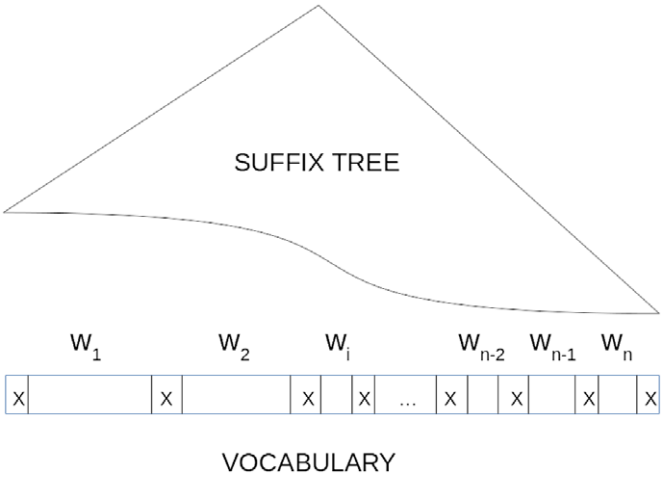

Figure 7. Vocabulary representation in our approximate matching system.

cases where the distance is below a threshold[ao]. An example of application of this rule is: *beta-adrenergic blocking agent* ← *beta adrenergic blocking agents*

- **Look for approximate shrink**: We look for the term candidate in the simple and complex STs. For all the fragments found, including a full word, we compute the string edit distance (Levenshtein distance) between these words and the candidate using the same thresholding mechanism. An example is *fluorouracil* ← *5-fluorouracil*.
- **Compatibility checking**: As an alternative to string similarity, we performed a compatibility checking between the term candidate and the terms in the ST. For this checking, a pre-process of lemmatisation, morphological analysis and semantic tagging has been performed using SpaCy[ap]. This checking consists of the sequential application of decreasing confidence compatibility rules, some of them requiring additional processes for their application, such as lemmatisation, WordNet linking and others. For instance, when the word forms are identical the compatibility is 1, when the two lemmas are identical the compatibility is 0.9 and when the difference is due just to the capitalisation there is an additional penalty of 0.05. Other rules check the WN overlap, the semantic distances in WN, the local hypernymy, hyponymy, and sibling relations in WN, etc. Some of the rules are language dependent. An example involving two variants of the same synset is *blocker* ⇔ *blocking agent*.
- **Removing diacritics**: In some cases, candidate terms lack diacritics and as the vocabularies contain diacritics the matching is more difficult or impossible. We have implemented a rule for facing this issue. An example in Spanish is *corazón* ⇐ *corazon (heart)*
- **Number variations**: When one of the terms to compare is singular and the other plural. An example is *antimuscarinic* ← *antimuscarinics*
- **Stop words removing in MW**: *sulfate atropine* ← *sulfate of atropine*.
- **Change of order in atomic components of a MW**: *h1 antagonist* ← *antagonist h1*.
- **Atomic components of a MW remotion.**: *h1 type antagonist* ← *h1 antagonist*.

The use of approximate matching substantially increases the coverage of the terminology for term look up but, obviously, the process is not error free as we discuss in Section 6.4.

[ao] We used 0.8 for hard approx. and 0.6 for loose approx.
[ap] https://spacy.io/

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

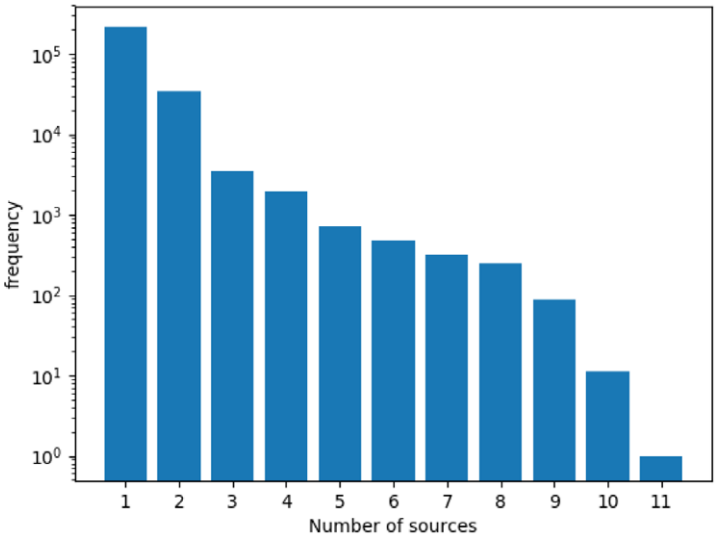

Figure 8. Histogram of sources per terms for English Body structure.

6.2 Intrinsic evaluation

Our terminology is built from a set of resources, as shown in Table 3, that support some degree of confidence on the two tasks involved, extracting the terminology and classifying each term into the appropriate category. As an individual term could be incorporated from more than one source, our intuition is that the number of sources (its support) is a good indicator of our confidence of this term for the corresponding category.

Consider, for instance, the case of the English language and 'Body structure' category. There are 256,179 terms in this vocabulary. The histogram of the number of sources per term is shown in Figure 8.

Although the shapes of the histograms for different languages and categories have some variations, the general distribution is similar to the one shown in Figure 8. In Table 11, we present the integrated (summing up all categories) results for all the languages. There is a good correlation between the distribution for English/Body structure and the global results in Table 11.

Although the distribution of sources is highly skewed towards terms coming from a unique source, the number of terms supported by more than one resource (almost 10%) assures a good confidence.

6.3 Extrinsic evaluation

To evaluate the content of a resource such as MHeTRep, we need one or more reference repositories or external material already validated. As the first possibility is not feasible, this evaluation was done against external resources (vocabularies, medical documents, contests material, etc.) for some languages and semantic categories.

Obviously, this evaluation covers only some of our cases. We have selected different sources to try to cover as many languages and semantic categories as possible. In most of cases, the sources correspond to learning and test material for NLP contests. In the late case, it is worth noting that we are not interested on the results of the task, but only on the list of medical terms validated by the organisation without discussing their validity or trying to do another term extraction task.

So, for each test we compute (1) the number of terms correctly extracted and classified using EXACT matching on *MHeTRep*, (2) the same using APPROX matching and (3) the coverage

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 11. Distribution of sources per term for all the terms languages and categories in MHeTRep

Number of sources	Terms	Percent
1	2,327,015	91.7
2	142,085	5.6
3	41,888	1.7
4	15,201	0.6
5	6082	0.2
6	2892	0.1
7	1431	0.1
8	737	0.0
9	208	0.0
10	34	0.0
11	9	0.0
12	2	0.0

using only the best individual resource. Our claim is that (3) should be overtaken by (1) and with great difference by (2). We have obtained the list of terms for each of the available resources and we have looked for these terms in *MHeTRep*. The search is done for EXACT matching and, in case of failure, for APPROX. The following sections briefly describe each of the sources used and assessments made.

6.3.1 DDI

The DDI Extraction 2013 task[aq] concerned the recognition of drugs and extraction of drug–drug interactions that appear in a corpus of biomedical English literature. Our test focuses on the first task. For this task, the DDI training corpus has been used (Herrero-Zazo *et al.* 2013).

We extracted all the terms tagged as 'Pharmaceutical/biologic product'. We collected 3638 terms of which 2896 are different. Looking at the presence of these terms within the English vocabulary for 'Pharmaceutical/biologic product', that is, looking for EXACT matching we obtained 1543 terms. These terms came from 11 different sources. The most frequent individual source was *dbpedia-en* contributing with 888 terms (30.7%), clearly under the global results. For the remaining terms, we further performed APPROX matching. 382 additional terms were found totalling 1925, that is, 66.5% of the candidate terms.

The conclusions from these results confirm our intuitions: (1) using a collection of terminologies improves the coverage of individual ones and (2) using APPROX matching allows to obtain an extra improvement in coverage.

6.3.2 MIMIC

MIMIC-III[ar] is a publicly available EHRs collection (Lee *et al.* 2011; Johnson *et al.* 2016), implemented as a database comprising deidentified health data associated with 40,000 critical care patients. For our test, we have selected only the tables containing descriptive fields:

[aq] https://www.cs.york.ac.uk/semeval-2013/task9.html.
[ar] https://mimic.physionet.org/

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 12. Result of MIMIC experiments using EXACT matching

#	Terms	Category	Exact terms	Alt cat	cat + alt	All/%	Source	Terms source/%
1	506	Subs	153/30	Pharm	261/52	348/69	DBP_en	104/21
2	10,675	Subs	345/3	Pharm	667/6	1143/11	Snomed-CT	232/2
3	4400	Subs	1038/24	Pharm	2070/47	2112/48	Snomed-CT	902/21
4	1193	CF	58/5	Pharm	78/7	92/8	Snomed-CT	28/2
5	23,274	CF	4722/20			4892/21	ICD-10	3194/14
6	6709	Proc	953/14			1112/17	Snomed-CT	912/14
All	46,757	–	7269/14	–	–	9699/21	–	5372/12

(1) D_ITEMS: Dictionary of local codes appearing in the MIMIC database.
(2) D_LABITEMS: Dictionary of local codes related to laboratory tests.
(3) DIAGNOSES_ICD: Diagnoses, coded using ICD[as].
(4) DRGCODES: Diagnosis Related Groups (DRG).
(5) PRESCRIPTIONS: Medications.
(6) PROCEDURES_ICD: Patient procedures, coded using ICD-PRM

Some extra comments are needed for this test. For each of the MIMIC tables, a category is expected. In some cases as 'D_ICD_DIAGNOSES' and 'D_ICD_PROCEDURES' what is expected is what is found (a 'Clinical finding' and a 'Procedure', respectively). In other cases, however, due to the different criteria used by MIMIC annotators and Snomed-CT ones the obtaining process is more problematic. To face this issue, we decided to detect the terms by EXACT matching in three alternative steps: (1) looking for the expected category, (2) looking for the pair of more likely categories and (3) looking for all the categories.

The results of our EXACT matching process are presented in Table 12. Column 1 shows the index of the MIMIC database table (shown above), column 2 its size (normalised unique terms), column 3 the expected category, column 5 the alternate category[at] and columns 4, 6, and 7 present the number of terms found for the three steps (expected category, first and second, all). We include the percent separated by a slash. Row 7 accounts for the overall results. Globally, MHeTRep contains 9699 terms from 46,757 existing in MIMIC (i.e., our coverage is 21%). Columns 8 and 9 present the best individual resource and its coverage. The most useful resource proved to be 'Snomed-CT' with an overall contribution of 5372 terms (12%) clearly above our results.

Then, we performed an APPROX matching search whose results are shown in Table 13. The overall results are that 62% of the pending terms are recognised, an excellent figure.

6.3.3 BARR acronym recognition contest

Acronyms have a high reference value, as they act as reference anchors of textual context most of the time. Because of this and the common problem of recognition of abbreviations, acronyms and

[as] MIMIC uses ICD-9 coding instead of ICD-10 used in our terminology. This is not relevant for our test.
[at] For 'D_ICD_DIAGNOSES' and 'D_ICD_PROCEDURES', as the expected category clearly outperforms the second one, the corresponding cell was left empty.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 13. Result of MIMIC experiments using APPROX matching

#	MIMIC table	Pending terms	App extracted	Pending
1	D_LABITEMS	158	88 (56)	70
2	D_ITEMS	9532	6847 (72)	2685
3	PRESCRIPTIONS	2288	1281 (56)	1007
4	DRGCODES	1101	973 (88)	128
5	D_ICD_DIAGNOSES	18,382	14,487 (79)	3895
6	D_ICD_PROCEDURES	5597	5427 (97)	170
–	ALL	37,058	29,103 (79)	7955

Table 14. Results obtained on BARR test

Type	Ann	EXACT matching	Best source	Best source size	APPROX matching	Both matching
short en	77	21	MCR_en	12	43	64
short es	311	37	Orphanet_es	20	170	207
long en	210	68	MCR_en	24	59	127
long es	930	166	MCR_es	66	260	426
total	1528	292 (19)		122 (8)	532 (35)	824 (54)

symbols, and their disambiguation[au], the Biomedical Abbreviation Recognition and Resolution (BARR) track was developed in 2017 (Intxaurrondo *et al.* 2017) within the framework of the Ibereval Campaign[av]. The task covered two languages, English and Spanish, and consisted on detecting short forms and long forms of acronyms or abbreviations occurring in EHRs. We have selected the learning data of this competition for testing our system. The recognition task is challenging, especially for recognising short forms.

The training data consisted of 525 documents containing 2959 annotations. Although the types of annotation consisted of 10 tags, we have reduced the cases to 'short' and 'long' annotations corresponding roughly to acronyms or abbreviations and their expansions. Results are presented in Table 14. Note that the 2959 annotations correspond to mentions (tokens) in the documents while the figures in column 2 of the table correspond to unique terms (types). Column 3 shows the figures of EXACT matching. The figures and the total percentage (19%) seem a bit low compared with the other tests, mainly due to the difficulty of the task for short forms. The distance between our full terminological resource (column 3) and the one for the best source (column 5) is maintained consistently.

The best source is MCR for the two English types and MCR and Orphanet for the Spanish case. The figures for APPROX matching are presented in column 6. The results are good, in fact the loss of EXACT match in this test is recovered by APPROX match. We have added a column combining the two matching for making clear this fact. Note that the results for short forms are consistently better than those for long forms for both languages (about 20%) while results for English are consistently better than those for Spanish for both short and long forms (about 15%).

[au]The same short form may correspond to several different long forms.
[av]http://nlp.uned.es/IberEval-2017/index.php

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 15. Results using the medical terms defined in the Clef eHealth competition

	#terms		Evaluation result			
			Full match			
Clef	Total	Unique	#/%	Main source		Partial match
2015	5237	2428	616/25.4	513/21.1	Snomed-CT	364/15
2016	5307	2462	671/27.3	574/23	Snomed-CT	390/15.8

6.3.4 Clef 2015 and 2016 contests

The CLEF Initiative (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum)[aw] focuses on multilingual and multi-modal information with various levels of structure.

CLEF eHealth offered challenges addressing several aspects of clinical information extraction including NER, normalisation and attribute extraction. Starting in 2015, the lab's IE challenge evolved to address lesser studied biomedical texts in languages other than English; in the 2015 and 2016 editions the language studied was French.

The dataset used in these competitions is the QUAERO French Medical Corpus. It was developed as a resource for NER and normalisation and fully described in Névéol *et al.* (2014). It includes a selection of MEDLINE titles and EMEA documents. The entities that were manually annotated are Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology, Procedures. All these categories were easily mapped into our tagset.

In Table 15, we present the results. Figures separated by a '/' represent the absolute number and the percentage in relation to the total. There is a high degree of correlation between all the figures for both years. Performing EXACT matching, we extracted a 25.4% (and 27.3%) of the terms. 364 (390) new terms, that is, 15% (15.8%) were located by APPROX matching. There is a high degree of correlation between the results obtained in both years. Performing EXACT matching, we extracted 25.4% (and 27.3%) of the terms. 364 (390) new terms, that is, 15% (15.8%) were located by APPROX matching. These figures clearly outperform the coverage obtained by the best source alone (cross-lingual links from English, Spanish and German).

6.3.5 Tass 2018

The Tass series of competitions are organised in the framework of the SEPLN[ax] conferences. Its aim is the furtherance of research in semantic tasks on texts written in Spanish in particular the processing of health data. eHealth-KD proposes the identification of two types of elements: concepts and actions. They can be linked by two types of relations: 'subject' and 'target' but we do not use these annotations.

The corpus used in this competition has been obtained from the MedlinePlus. It brings information about diseases, conditions and wellness issues in understandable language for everybody. For this contest, only Spanish language items were considered. Once cleaned, each selected item was converted to a plain text document and pre-processed to remove unwanted sentences (headers, footers and similar elements) and to flatten HTML lists into plain sentences. The final documents are manually tagged using Brat.

[aw] http://www.clef-initiative.eu/
[ax] http://www.sepln.org/

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 16. Results using the medical terms collected from medical discharge reports

	#terms		Evaluation result			
			Full match			
EHR	Total	Unique	#/%	Main source		Partial match
25%	6137	35	19/54	16/46	Snomed-CT	3/8.6
50%	12,238	193	103/53	78/40	Snomed-CT	32/16.6
75%	18,346	851	364/43	276/32	Snomed-CT	170/20
100%	24,465	4817	1113/23	868/18	Snomed-CT	936/19.4

Performing EXACT matching, we extracted 336, that is a 34%, of the unique terms. 122 new terms (12%) were located by APPROX matching. These figures clearly outperform the coverage obtained by the best source alone (cross-lingual links from French DBP) that extracted 139 terms, that is, 13% of the terms.

6.3.6 Structured discharge data from HIBA

This set of terms has been compiled from a set of 2819 HIBA[ay] discharge notes. HIBA is a reference hospital in Buenos Aires closely involved in Snomed-CT implementation.

The results are presented in Table 16. From 4827 unique terms, 1113 have been recognised by EXACT matching (23%). This figure is in line with the corresponding figures in other datasets. From these 1113 terms, 868 (18%) come from one source: Snomed-CT. If we do a stratified selection, we can see that only 35 terms constitute 25% of all terms present in the collection. From this quantity, more than 60 are found in *MHeTRep*. These figures are relatively high. In fact, the terms come from 11 different contributors, most of them coming from 'Snomed-CT' (the second contributor, 'DBPedia_es', provides only 439 terms). The APPROX matching procedure allows the detection of 936 terms. The combination of these facts provides a reasonable explanation of the figures. APPROX and EXACT match contribute similarly, coming both mainly from the 'Snomed-CT'. So, the conclusion is that the annotators of the documents seem to accurately follow the 'Snomed-CT' terminology.

6.3.7 SemEval 2017 task 3 subtask D (Arabic community question answering)

This contest refers to Community Question Answering on medical fora for Arabic. Although Arabic has a relatively poor coverage in our terminologies, we think that it can be useful to perform a test on this language also. Additionally, the data for this task come from a challenging genre, medical fora. The SemEval Task 3 subtask D, (Nakov *et al.* 2017), asks, given a query, consisting of a question in Arabic, and a set of question–answer pairs (for each query, 30 for learning, 9 for testing), (1) to classify each pair as relevant or not regarding the query and (2) to re-rank the question–answer pairs according to their relevance with respect to the original question. An important issue we have to face is that this material lacks annotations of medical entities.

Thus, for selecting the candidate terms, we extracted all the word n-grams (up to 5-grams) satisfying termhood conditions, that is, combinations[az] as N, NN, JN, . . ., for English or N, NN,

[ay] Hospital Italiano de Buenos Aires.
[az] The meaning of each letter is the following: N = noun, J = adjective.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 17. Results on SemEval 2017 test (BLE: Bilingual extension, BC: Best count)

Lang	Snomed categ.	Cand	Terms	Exact	Approx	E+A	Best	BC
en	SUB	230,433	11,291	470	722	1192	Snomed-CT	252
en	CLF	453,543	22,224	497	616	1113	BLE_fr_en	279
en	BOS	255,703	12,529	396	617	1013	FMA	189
en	PHA	157,159	7701	439	566	1005	Galen	189
en	PRO	289,497	14,185	115	189	304	MCR_en	60
en	ORG	75,200	3685	73	272	345	BLE_fr_en	41
en	Total	1,461,535	71,615	1990	2982	4972		1010
ar	SUB	5406	860	47	224	271	BLE_fr_ar	47
ar	CLF	3866	615	172	234	406	BLE_en_ar	153
ar	BOS	3118	496	81	242	323	BLE_en_ar	74
ar	PHA	4464	710	84	226	310	BLE_fr_ar	72
ar	PRO	547	87	22	64	86	BLE_en_ar	28
ar	ORG	2244	357	35	35	70	BLE_en_ar	17
ar	Total	19,645	3124	441	1025	1466		391

NJ, . . ., for Arabic. Note that the usual approaches to this task are based on computing the semantic similarity between the query and the query and answer of the thread pairs. Locating the medical tokens in these strings is the core task of these approaches.

As we have of the Arabic texts for all the questions, and questions and answers for all the pairs as well as the English translations using Google Translate API[ba] we will test the coverage for both languages in order to compare the results. Results are presented for both languages and semantic classes considered in Table 17.

The third column of Table 17 reflects the sizes of the word n-grams vocabularies. These vocabularies, however, contain many out-of-domain n-grams. In order to measure the degree of coverage of our terminologies, we have estimated the proportion of true medical terms within the sets of candidates. We have carried out this estimation by manually examining a set of 700 sentences (for each language) from our corpus. The sizes of these estimations are presented in column 4 of the table.

Next columns present the number of exact matches, approximate matches and overall matches for each semantic class and language. The best source and its results are shown in columns 8 and 9.

It is worth noting that the best sources for Arabic correspond to DBPs for English and French. The results for Arabic, 441 exact matches (14.1%), 1025 for approximate matches (32.8%), totalling 1466 (46.9%), are similar to the obtained in other tests. However, the results for English, are very low. The reason for this is probably due to the bad quality of the translations of the Arabic original texts, especially for some semantic classes.

In order to evaluate the use of *MHeTRep* in SemEval Task 3 subdues D, we have slightly modified our algorithm by including the Arabic vocabularies of *MHeTRep*. The original algorithm is described in Adlouni *et al.* (2017)[bb]. In Table 18, we present the official result of the primary submissions (three first rows).

[ba] https://cloud.google.com/translate/docs/

[bb] The system is a joint collaboration between UPC and USMBA.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 18. Official results of the task and improved version

Team	Rank	MAP	Accuracy
GW_QA-primary	1	0.6116	0.6077
UPC-USMBA-primary	2	0.5773	0.6624
QU_BIGIR-primary	3	0.56695	0.4964
UPC-USMBA-improved		0.5913	0.6815
Improvement (%)		2.43	2.88

For improving the results using MHeTRep, we have looked for the occurrences of terms in our vocabularies in each of the 1400 queries in the test dataset. From the 1400 queries, 1239 include at least one term of *MHeTRep*. For these queries, we looked for all the recognised terms into the question/answer pairs in the thread. 6154 matching were obtained (on average 5 terms per query). We have added a new feature to our algorithm for learning a classifier consisting on the size of the intersections of the sets of *MHeTRep* terms in the original query and the candidate pair. After running the updated algorithm, we obtained an improvement in two metrics (MAP, 2.43%, accuracy, 2.88%).

6.4 Discussion on extrinsic evaluation

From the evaluation presented above, some general considerations can be done:

(1) MHeTRep has been tested using mainly sources where the terminological strings have been already identified (all but the SemEval 2017 cases). Such sources have been compiled from different origins including medical articles, discharge notes, EHRs and contest materials. Some issues can be identified:

   — In some cases, the reliability is not guaranteed as they has been tagged by the contest organiser. Data obtained from Tass 2018, for example, has been obtained from text with a low specialisation level (e.g. *enfermedad* -disease-, *resultado* -result-, . . .). Some of the validated terms may not even be considered medical terms (e.g. *proveedor* -provider-, *insecticidas* -insecticides-, *pintura* -paint-, . . .).
   — Sometimes the terms contain spelling errors due to the process of building. See for example: *sospecha de sepsis* or *úlceras por presion en el talon*; such strings contains extra words or spelling errors making its identification impossible or they can even contain more than one term.

(2) Most of the resources used to compile our collection include just specific items, not their super-classes. For example, in evaluating DDI it is possible to observe that most of the unrecognised terms correspond to generic drugs (e.g. *adrenergic bronchodilators*, *alpha glucosidase inhibitors* and *ergot alkaloid class* among many others). Obviously, they are chemical compounds widely used in medicine but the resources analysed to build our term collection only include specific compounds not their super-classes. A similar issue appears with the disease names, if such a name correspond to a group of diseases it can happen that the generic name is not included in such a resource. See, for example, *cancers digestifs* or *maladies du foie* in French and *infecciones urinarias* and *problemas de salud* in Spanish.

(3) We also used the *WP* as a resource to build our collection but it causes a problem: it uses a redirection mechanism to direct the user to some kind of standard terms. But it happens that this mechanism is ambiguous because it is also used to correct some user spelling

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 19. Examples of approximate matching results. (all are English terms come from DDI, examples 11 and 12 is a Spanish example and example 13 in a French example

#	Term to evaluate	Actual term	Result	Coef.
1	vitamin d 3	vitamin b3	ko	0.82
2	vitamin d 3	vitamin d3	ok	0.91
3	articaine hydrochloride	ranitidine hydrochloride	ko	0.83
4	penicillamine	d-penicillamine	sub	0.87
5	cyclosporins	cyclosporin a	sub	0.85
6	dermax	permax	ko	0.83
7	h1 antagonist	h2 antagonist	ko	0.86
8	$\beta - lactam$ antibiotic	b –lactam antibiotic	ok	0.90
9	non-opioid analgesics	opioid analgesics	ko	0.81
10	potassium-sparing diuretics	non-potassium sparing diuretic	ko	0.84
11	nacimiento bebé	see text	ko	?
12	número	húmero	ko	0.83
13	coleur	douleur	ko	0.71

mistakes (ex *rinon* redirects to *riñon* -kidney-). It is also used to resolve ambiguities (remember the typical "bank" example). For this reason, we do not use this information to collect terms; therefore a string like *fat-soluble vitamin* that is redirected to *vitamin* is not included in our collection.

Regarding our approximate matching algorithm, it should be noted that although this technique is useful for recognising terms that have not been correctly spelled, but some nonrecognition can occur. The algorithm works correctly when it meets terms where an accent is missing, the term is in the plural or there are simple errors in the order of the characters. But we cannot understand that all errors are corrected; it may happen that misrecognition occurs; that is, the recognised term may have a dubious relationship with the original term. Table 19 shows some examples of these situations.

2 and 8 are examples of misspelled terms although our algorithm successfully solves them. The first needs only one insertion and reaches a score 0.91. The second has a score 0.9 and requires one insertion and one modification but the length is higher.

4 and 5 correspond to misdetection: The term under evaluation is a direct hypernym of the term saved in our collection. It requires the removal of *d* and *a* and two deletions.

1 and 7 are errors probably because the right terms are not included in the vocabularies. The term *vitamin d 3* exists in the vocabulary, *vitamin b3* does not exist but it should be included. The same observation can be applied to *h1* and *h2* as modifiers of antagonist.

9 and 10 are examples of negated terms. In 9 the negated term exists in the vocabulary but not the affirmed one. In 10, the situation is the contrary. When the negated term occurs in the vocabulary the affirmed term should occur as well and has to be included. When only the affirmed term occurs the inclusion of the negated term is dubious, the negation of a term probably is not a valid term, this situation should be carefully analysed by a terminologist. 3 and 6 correspond to cases of terms corresponding to words that are similar to other terms included in our vocabulary. As a matter of fact, *dermax* correspond to just a trademark of a therapeutic shampoo that has been wrongly tagged as a term.

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 20. Coverage of Snomed-CT classes of English terms from different sources

row	Source	BOS	CLF	EVT	OEN	ORG	PHA	PHO	PRO	QUV	SUB	Other
1	Biling	3949	4013	787	0	3482	6485	0	821	0	7156	0
2	DBP	36,504	16,554	0	0	0	57,231	0	0	0	90,214	0
3	DB	0	0	0	0	0	17,539	0	0	0	0	0
4	FMA	176,491	0	0	0	0	0	0	0	0	0	0
5	GAL	8159	8916	42	69	58	29,448	64	217	236	237	61
6	ICD	0	173,101	0	0	0	0	0	173,415	0	0	0
7	MCR	4161	2745	1449	190	0	1116	769	1005	529	4259	115
8	MeSH	1707	8758	263	2472	2787	47,390	358	3148	1645	139,348	1071
9	OPHM	0	42,972	0	0	0	0	0	0	0	0	0
10	RAD	20,265	785	0	1	95	95	240	0	23	0	0
11	SN	55,755	215,733	7230	17,356	70,664	36,188	30,166	111,762	18,808	51,784	27,803
12	Other	5321	12,082	0	0	0	12,935	0	0	0	0	0
13	max	176,491	215,733	7230	17,356	70,664	57,231	30,166	173,415	18,808	139,348	27,803
14	Δ_m	0.43	0.56	0.26	0.14	0.08	0.73	0.04	0.4	0.11	0.52	0.04
15	*res_m*	FMA	SN	SN	SN	SN	DBP	SN	ICD	SN	MeSH	SN
16	Δ_SN	0.82	0.56	0.26	0.14	0.08	D0.83	0.04	0.62	0.11	0.82	0.04

In example 3, *articaine hydrochloride* corresponds to an anaesthetics not included in our vocabulary but whose spelling is similar to a chemical compound included in the vocabulary with a quite different purpose.

11 corresponds to a term not found (*nacimiento bebé*) in spite of the fact that very similar terms are included (*nacimiento de un niño*) in Snomed-CT and may be considered as a variant of the same concept. In this case, it seems clear that the terminology used in this institution do not exactly match the one proposed by Snomed-CT.

Finally, 12 and 13 reaches a score of 0.83 and 0.71 but even though they require 1 and 2 single character replacements they are clear mistakes. This is a particularly complicate case since the word that the algorithm manipulates leads to a word that corresponds to a term that has no relation to the origin. These are not real examples, they has been included just to show the difficulty of the task.

Some of the issues mentioned above (like negation) could be solved relatively easily by improving our algorithm but probably its solution would be language dependant. Some issues requires a treatment specific for those terms that includes a number (i.e., *vitamin d 3* vs *vitamin b 3* or *hepatitis type I* vs *hepatitis type II*). Other cases requires an enlargement of the vocabulary.

Table 21 shows a full view of all the evaluations. In spite of the issues regarding the evaluation material mentioned above, the results of the evaluation are acceptable. A remarkable exception is the evaluation of terms in the SemEval competition due to requirement to use automatic translation of Arabic texts.

It is important to check to what extent using *MHeTRep* vocabularies is better than using directly one or more of the original sources. We consider two cases depending of the language: languages other than English and English. In the first case the analysis is clear, the coverage is

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Table 21. Summary of the results obtained by extrinsic evaluation

Competition	Lang	#	Text type	Full match	Partial match	Total
DDI	EN	3628	articles	53.3	13.2	60.5
MIMIC	EN	46,757	EHR	12.0	62.0	74.0
Acronyms Comp.	EN	1528	acronyms	19.0	35-0	54.0
Clef 2015	FR	5237	articles	25.4	15.0	40.4
Clef 2016	FR	5307	articles	27.3	15.8	43.1
Tass	ES	1805	articles	34.0	12.0	46.0
Discharge data	ES	6137	EHR	54.0	8.6	62.6
SemEval	EN	1,461,535	fora	2.8	4.2	6.9
	AR	19,645	fora	14.1	32.8	46.9

heavily dependent of the terms obtained from other languages, mainly English, due to the limited amount of its own resources. Such languages range from Spanish, with a relatively high number of resources to Arabic with a 100% of their terms coming from other languages. In any case, the number of resources and their completeness is much lower than for English.

For English, we need a more accurate analysis. In Table 20, rows 1 to 12 show the contribution of the different sources to the vocabularies of the different classes in columns. As expected, the contribution of each source alone is lower than the global coverage of *MHeTRep* but let us look more in depth. In rows 15 and 13, we present the first contributor for each class and the number of terms contributed. For most of the classes, the first source was *Snomed-CT* but for some important ones the first contributor was another source (*FMA*, *DBPedia*, *MeSH)*. We devise, thus, two scenarios: (i) using only one resource, that is, *Snomed-CT*, and (ii) using the best resource for each class. We computed, for both cases and for each class the fraction of the *MHeTRep* vocabulary covered by the corresponding source. Δ_SN (row 16) covers the former case and Δ_m (row 13) the latter. For instance, for *Body Structure* (column 2), the best contributor is *FMA* with 176,491 terms. As the size of *MHeTRep* for this class is 312,312 terms, 43% of the terms (row 14) are not covered by the best resource. In the second case, using *Snomed-CT*, this resource covers only 55,755 and, so, 82% of the terms are not provided by this resource. The rest of classes confirm this figures and, so, we can conclude that *MHeTRep* performs the best in both scenarios.

7. Conclusions

In this paper we have proposed, applied, and evaluated a methodology for building *MHeTRep*, a semantically tagged multilingual terminology from available lexical resources in the medical domain. Our main contribution is a rich terminology comprising about 2,5 terms that we have made public.

Our claim is that the methodology proposed is partially language independent. Only shallow linguistic processors are needed, and the application to whatever language only depends on the availability of medical lexical resources for such language.

We have applied the methodology to seven languages. As lexical resources, we have used 13 datasets, which are presented in Table 3.

Among these resources, there are monolingual (as MeSH, Galen, FMA, for English, or CIMA for Spanish), multilingual (as Snomed-CT, for Catalan, English, and Spanish, Orphanet, for English, French, German, and Spanish, or ICD-10 for English, French and Spanish) and cross-lingual resources (as the different versions of WN and the DBPedia). As a semantic tagset,

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

we have used the top-level categories defined in Snomed-CT (see Table 2). For building the terminology, we proceed into two steps: (1) a monolingual phase applied independently for each resource and eventually for each language and semantic category and (2) a multilingual extension using cross-lingual resources that is applied iteratively until no further enrichment is obtained. The second phase is applied iteratively until no further enrichment is obtained.

For each pair <language, resource>, we proceed in two steps: (1) collection of the terms from the resource and (2) classification into the appropriate category. The compilation process is usually straightforward, except for ontologies that merge conceptual and lexical entities. In general, the classification step is more challenging. In some cases, the classification step is simple (for instance, terms in DO are always classified as 'Clinical finding', etc.) but in general presents more serious issues.

We defined two evaluation scenarios: an intrinsic evaluation and an extrinsic one. The intrinsic scenario basically evaluates the classification subtask and is based on the idea that higher the number of resources that agree on a specific category the higher the confidence in such a classification. Although most of the collected terms come from a unique resource, there is a non-negligible percentage (about 10% on average) of terms coming from more than one source. This figure ensures a good level of confidence on our task.

We complemented the intrinsic evaluation with a set of opportunistic extrinsic (indirect) evaluation. For this task, we selected cases in which the material already annotated was already available trying to cover as many languages and semantic classes as possible. We measured the ratio of coverage of the annotated terms by our terminology using full and approximate matching. The overall results are presented in Table 9. Summarising plots are shown in Figures 3 and 4. The result shows that MHeTRep consistently outperforms the best only one resource by 20% for all the languages and cases.

Regarding the coverage per language, as expected, the richest terminology for all the categories was the English one, with an overall size of more than 1.5 M terms. At a considerable distance, we find the Spanish terminology with about 0.64 M terms. All the other languages show smaller coverage. The language with the lowest coverage is Arabic due to the lack of specific medical resources.

We conclude by stating that the resource proposed in this paper can be useful whenever a terminology of the medical domain is needed in a certain language and it is necessary to cover either the entire domain or a specific semantic field.

8. Future work

As future work, we plan the following actions:

- To extend the coverage of MHeTRep and the accuracy of the classification processes.
- Including new resources available for less resourced languages.
- Extending English coverage with resources covering sub-domains. A more systematic use of the huge material contained into the BioPortal ontologies is a basic starting point in this line.
- Using annotated material (the datasets used for extrinsic evaluation and other similar) for enriching the terminologies.
- Apply the methodology to other languages in order to extend heavily the number of covered languages, focusing on less resourced languages and those owning non-Latin scripts, as Asiatic and Slavic languages. Multilingual resources as *BabelNet*[bc] (Navigli and Ponzetto

[bc] https://babelnet.org/

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

2012) or *Open Multilingual Wordnet*[bd] Bond and Foster (2013) should be used for this purpose.

- Improve our approximate matching approach analysing the cases of failure sketched in Section 6.4.

Acknowledgements. The author Jorge Vivaldi was partially funded by the public supported project TERMMED (FFI2017-88100-P, MINECO). The author Horacio Rodríguez was partially supported by the public funded project GRAPHMED (TIN2016-77820-C3-3R).

Furthermore, the authors are grateful for any suggestions and advice received from the anonymous reviewers. Your collaboration has allowed us to improve this paper.

References

Aronson A. and **Lang F.** (2010). An overview of MetaMap: Historical perspective. *Journal of the American Medical Informatics Association* **17**, 229–236.

Athenikos S. J. and **Han H.** (2010). Biomedical question answering: A survey. *International Committee on Computational Linguistics* **99**(1), 1–24.

Batet M. and **Sánchez, D.** (2020). Leveraging synonymy and polysemy to improve semantic similarity assessments based on intrinsic information content. *Artificial Intelligence Revue* **53**(3), 2023–2041.

Bay M., **Bruneÿ D.**, **Herold M.**, **Schulze C.**, **Guckert M.** and **Minor M.** (2021). Term extraction from medical documents using word embeddings. *Proceedings of 2020 6th IEEE Congress on Information Science and Technology (CiSt)*, pp. 328–333.

Bodenreider O. (2004). The Unified Medical Language System (UMLS): Integrating biomedic terminology. *Nucleic Acids Research* **32**, 267–270.

Bond F. and **Foster R.** (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL*, 4-9 August 2013, Sofia, Bulgaria. pp. 1352–1362.

Cabré M., **Estopà R.** and **Vivaldi J.** (2001). Automatic term detection: A review of current systems. In **Bourigault D.**, **Jacquemin C.** and **L'Homme M. C.** (eds), *Recent Advances in Computational Terminology*. John Benjamins. pp. 53–87.

Cohen K., **Verspoor K.**, **Fort K.**, **Funk C.**, **Bada M.**, **Palmer M.** and **Hunter L.** (2017). The Colorado Richly Annotated Full Text (CRAFT) corpus: Multi-model annotation in the biomedical domain. *Handbook of Linguistic Annotation*, pp. 1379–1394.

Corsar D., **Moss L.**, **Sleeman D.** and **Sim M.** (2009). Supporting the development of medical ontologies. *Proceedings of the 4th Workshop Formal Ontologies Meet Industry*, pp. 114–125.

Conrado M., **Pardo T.** and **Rezende S.** (2013). *Exploration of a Rich Feature Set for Automatic Term Extraction. Lecture Notes in Computer Science*, Vol. **8265**, Berlin Heidelberg: Springer, pp. 342–354.

Dinh D. and **Tamine L.** (2011). Voting techniques for a multi-terminology based biomedical information retrieval. In **Peleg M.**, **LavraČ N.** and **Combi C.** (eds), *Artificial Intelligence In Medicine, Aime 2011*. Lecture Notes in Computer Science, Vol. **6747**. Berlin, Heidelberg: Springer.

El Adlouni Y., **Lahbari I.**, **Rodriguez H.**, **Meknassi M.**, **El Alaoui S.** and **Noureddine N.** (2017). UPC-USMBA at SemEval-2017 Task 3: Combining multiple approaches for CQA for Arabic. doi= 10.18653/v1/S17-2044. pp. 275–79.

de Freitas F., **Schulz S.** and **Moraes E.** (2009). Survey of current terminologies and ontologies in biology and medicine. *Electronic Journal in Communication, Information and Innovation in Health*, 7–18.

Donnelly K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in Health Technology and Informatics* **121**, 279–290.

Drouin P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology* **9**, 99–115.

Enguehard, C. and **Pantera, L.** (1995). Automatic natural acquisition of a terminology. *Journal Of Quantitative Linguistics* **2**, 27–32.

Frantzi K., **Ananiadou S.** and **Tsujii J.** (1998). The C-value/NC-Value Method of Automatic Recognition for Multi-word Terms. Berlin Heidelberg: Springer, pp. 585–604.

Gomaa W. H. and **Fahmy A. A.**, et al. (2013). A survey of text similarity approaches. *International Journal of Computer Applications* **68**(13), 13–18.

Gonzalez-Agirre A., **Laparra E.** and **Rigau G.** (2012). *Multilingual Central Repository version 3.0*. European Languages Resources Association (ELRA), pp. 2525–2529.

Goodwin T., **Savery M.** and **Demner-Fushman D.** (2020). Flight of the PEGASUS? Comparing transformers on few-shot and zero-shot multi-document abstractive summarization. *Computer Methods and Programs in Biomedicine*, pp. 640–5646.

[bd] http://compling.hss.ntu.edu.sg/omw/

https://doi.org/10.1017/S1351324922000055 Published online by Cambridge University Press

Heid U., **Jauss S.**, **Krueger K.** and **Hohmann A.** (1996). Term extraction with standard tools for corpus exploration. Experience from German. In *Proceedings of TKE'96 Terminology and Knowledge Engineering*, pp. 139–50.

Hellrich J., **Schulz S.**, **Buechel S.** and **Hahn U.** (2015). JUFIT: A configurable rule engine for filtering and generating new multilingual UMLS terms. In *American Medical Informatics Association Annual Symposium*, pp. 604–610.

Herrero-Zazo M., **Segura-Bedmar I.**, **Martínez P.** and **Declerck T.** (2013). The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics* **46**, 914–920.

Intxaurrondo A., **Pérez-Pérez M.**, **Pérez-Rodríguez G.**, **López-Martín J.**, **Santamaria J.**, **de la Peña S.**, **Villegas M.**, **Akhondi S.**, **Valencia A.**, **Lourenço A.** and **Krallinger M.** (2017). The Biomedical Abbreviation Recognition and Resolution (BARR) track: benchmarking, evaluation and importance of abbreviation recognition systems applied to Spanish biomedical abstracts. In *Proceedings of SEPLN 2017*, pp. 230–246.

Johnson A., **Pollard T.**, **Shen L.**, **Lehman L.**, **Feng M.**, **Ghassemi M.**, **Moody B.**, **Szolovits P.**, **Celi L.** and **Mark R.** (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data.* doi: 10.1038/sdata.2016.35.

Jonquet C., **Emonet V.** and **Musen M.A.** (2015). Roadmap for a Multilingual BioPortal. In *Proceedings of the Fourth Workshop on the Multilingual Semantic Web (MSW4) co-located with 12th Extended Semantic Web Conference (ESWC)*, pp. 15–26.

Lamy J., **Venot A.** and **Duclos C.** (2006). PyMedTermino: an open-source generic API for advanced terminology services. *Studies in Health Technology and Informatics* **210**, 924–928.

Lastra-Díaz J.J., **Goikoetxea J.**, **Ali Hadj Taieb M.**, **García-Serrano A.**, **Ben Aouicha M.**, **Agirre E.** and **Sánchez D.** (2021). A large reproducible benchmark of ontology-based methods and word embeddings for word similarity. *Information Systems* **96**(3), 1016–1036.

Lipscomb C. (2006). Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association* **88**(3), pp. 265–266.

Lee J., **Scott D.**, **Villarroel M.**, **Clifford G.**, **Saeed M.** and **Mar R.** (2011). Open access MIMIC-II database for intensive care research. In *33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 8315–8318.

Lossio-Ventura J., **Jonquet C.**, **Roche M.** and **Teisseire M.** (2016). Biomedical term extraction: Overview and a new methodology. *Information Retrieval* **19**, 59–99.

Mejino J. Jr., **Agoncillo A.**, **Rickard K.** and **Rosse C.** (2003). Representing complexity in part-whole relationships within the foundational model of anatomy. In *AMIA Symposium*, vol. 2003, pp. 450–454.

Mozzicato P. (2009). An overview of the medical dictionary for regulatory activities. *Pharmaceutical Medicine* **23**, 65–75.

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

Nakov P., **Hoogeveen D.**, **Màrquez L.**, **Moschitti A.**, **Mubarak H.**, **Baldwin T.** and **Verspoor K.** (2017). *SemEval-2017 Task 3: Community question answering*. Association for Computational Linguistics, pp. 27–48.

Navigli R. and **Ponzetto S.** (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* **193**, pp. 217–250.

Newman D., **Koilada N.**, **Lau J.** and **Baldwin T.** (2012). Bayesian text segmentation for index term identification and Keyphrase extraction. In *Proceedings of COLING 2012*, p. 2077–92.

Névéol A., **Dalianis H.**, **Velupillai S.**, **Savova G.** and **Zweigenbaum P.** (2018). Clinical natural language processing in languages other than English: Opportunities and challenges. *Journal of Biomedical Semantics* **9**(1), 12.

Noy F.N., **Musen M.A.**, **Mejino J.** and **Rosse C.** (2004). Pushing the envelope: Challenges in a frame-based representation of human anatomy. *Data Knowledge Engineering* **48**, 335–359.

Pazienza M., **Pennacchiotti M.** and **Zanzotto F.** (2005). *Terminology Extraction: An Analysis of Linguistic and Statistical Approaches*. Berlin, Heidelberg: Springer, pp. 255–279.

Rector A., **Qamar R.** and **Marley T.** (2009). Binding ontologies and coding systems to electronic health records and messages. *Applied Ontology* **4**, 51–69.

Rosse C. and **Mejino J.** (2003). A reference ontology for biomedical informatics: The foundational model of anatomy. *Journal of Biomedical Informatics* **36**, 478–500.

Rubin, **D.** (2008). Creating and curating a terminology for radiology: Ontology modeling and analysis. *Journal of Digital Imaging* **21**, 355–362.

Smith B. and **Scheuermann R.** (2011). Ontologies for clinical and translational research: Introduction. *Journal of Biomedical Informatics* **44**, 3–7.

Salvadores M., **Horridge M.**, **Alex P.**, **Fergerson R.**, **Musen M.** and **Noy N. F.** (2012). Using SPARQL to query BioPortal ontologies and metadata. In *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference*, pp. 180–195.

Shickel B., **Tighe P.**, **Bihorac A.** and **Rashidi P.** (2018). Deep EHR: A survey of recent advances in deep learning techniques for Electronic Health Record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics* **22**, 1589–1604.

Wishart D., **Knox C.**, **Guo A.**, **Shrivastava S.**, **Hassanali M.**, **Stothard P.**, **Chang Z.** and **Woolsey J.** (2006). DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* **34**, D668–D672.

Wishart D., **Feunang Y.**, **Guo A.**, **Lo E.**, **Marcu A.**, **Grant J.**, **Sajed T.**, **Johnson D.**, **Li C.** and **Sayeeda Z.** (2017). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082.

Wang R. and **Liu W.** (2016). Featureless domain-specific term extraction with minimal labelled data. In *Proceedings of Australasian Language Technology Association Workshop*, pp. 103–112.

Whetzel P. L., **Noy N. F.**, **Shah N. H.**, **Alexander P. R.**, **Nyulas C.**, **Tudorache T.** and **Musen M. A.** (2011). BioPortal: Enhanced functionality via new web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research* **39**, 541–545.

Young T., **Hazarika D.**, **Poria S.** and **Cambria E.** (2017). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* **13**(2), 55–75.

Cite this article: Vivaldi J and Rodríguez H (2023). MHeTRep: A multilingual semantically tagged health terms repository. *Natural Language Engineering* **29**, 1364–1401. https://doi.org/10.1017/S1351324922000055