Automatically Producing Semantically Tagged Bilingual Terminologies

Jorge Vivaldi¹ · Horacio Rodríguez²

Received: 21 October 2020 / Accepted: 23 October 2021 / Published online: 19 November 2021
© The Author(s) 2021

Abstract

Even though many *NLP* resources and tools claim to be domain independent, their application to specific tasks is restricted to some specific domain, 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. This paper proposes a method for automatically compile terminologies from potentially any domain. The proposed method takes as reference the set of domains defined by Magnini, the Multilingual Central Repository (a resource based on *WordNet* 3.0) together with *DBpedia*, an open knowledge source that had proven to be reliable for restricted domains. Using the method described in this article, we have produced a big set of reliable terminologies for 164 domains and 2 languages totalling 635,527 terms. The proposed method has been applied to English and Spanish languages but it is potentially applicable to any language that has its own a *DBpedia* evolved enough. The obtained results have been intensively evaluated in several ways.

Keywords Multi-domain term collection and bilingual terminologies · MCR-based terminologies

Introduction

Compiling terminologies for a domain of interest is an initial step needed for whatever *NLP* task in such domain. Domain adaptation of existing resources (data or processors) for performing tasks as semantic tagging, relation extraction, semantic parsing, semantic role labelling, etc. within domain restricted applications, as Question Answering, Automatic Summarization, Machine Translation, etc. is heavily dependent of disposing of such terminologies. Also the emerging discipline of cross-lingual transfer learning [39] can get improvements using this kind of resources.

Terminology Extraction, *TE*, is, however, a difficult task, and performing *TE* manually is highly time consuming and prone to errors (sometimes false positives but mainly false negatives), especially in the case of multiword terms. The difficulty applies to both the termhood of the candidate as well its scope. Therefore, performing the task automatically is practically mandatory. There are, fortunately, for some

languages, as English, some lexical resources where these terms are collected and the problem is to locate these term candidates, *tc*, in the text and classify them as belonging or not to the terminology of the domain.

Our aim is, given (1) a semantic tagset of domains, likely covering the whole semantic space, (2) a pair of languages, and (3) a set of lexical sources, covering all the domains of the tagset for the two languages, collecting from the lexical sources for each domain in the tagset and for the two languages as many terms in the domain as possible and trying to map the corresponding terms between the two languages.

Our main objective is getting accurate and, if possible, huge terminologies for each domain for the two languages involved. Obtaining bilingual mappings is a simple lateral effect of our method. Terms are mapped only in the case an explicit link exists in any of the resources used. No attempt has been made to apply automatic mapping procedures, never error free.

In Table 1, we present the notations most frequently used along the paper.¹

The process for each source, or combination of sources, is threefold:

✉ Jorge Vivaldi
jorge.vivaldi@upf.edu
Horacio Rodríguez
horacio@lsi.upc.edu

¹ Pompeu Fabra University, Barcelona, Spain

² Polytechnical University of Catalonia, Barcelona, Spain

¹ *l* and *s* are used as generic tags for language and source. When referring to specific ones, we can instantiate them to *en* or *es* and *wp*, *dbp* or *wn* respectively.

Table 1 Notation used in the paper

Term	Definition
WP^l	Wikipedia for language l
WP_{CG}^l	Wikipedia category graph for language l
WP_{PG}^l	Wikipedia page graph for language l
WN^l	WordNet for language l
WND	WordNet domain codes
$XWND$	Extended WordNet domain
EWN	EuroWordNet
DBP^l	DBpedia for language l
WP_{pages}^l	Wikipedia graph of pages for language l
$WP_{categories}^l$	Wikipedia graph of categories for language l
t	Term
tc	Term candidate
tcS	Set of term candidates
l	language
s	Lexical source
db	Domain border
d	Domain
dc	WN domain code, $dc \in WND$
$page$	Wikipedia page
syn	WordNet synset
$tcS_{l,d}^s$	Set of term candidates for language l , source s , and domain d
$tS_{l,d}$	Set of terms for language l and domain d
$eval_{emb}^{l,dc}$	Set of terms for evaluating embeddings for the domain dc and the language l .

- Given a domain d , a language, l , and a lexical source s , likely having a high density of terms from d , extract from s a set of term candidates (tcS).
- Filter tcS for removing from it those tc not according to termhood conditions.²
- Filter out tcS not belonging to the domain d .

In this task, we use: (1) as semantic tagset Magnini's Wordnet Domains Codes, WND , as discussed in “[WordNet domains](#)”, (2) the following lexical sources: Wikipedia³ (WP), *DBpedia*⁴ (DBP), WordNet⁵ (WN), [12, 13] for two

languages, English and Spanish. The extracted terms are restricted to be WP pages or category titles.

Our claim is that our approach could be used for other languages disposing of WP , DBP , and WN and reliable resources for doing POS tagging.⁶

Let $tcS_{l,d}^s$ be the set of term candidates for domain d , language, l , and lexical source s . For instance, $tcS_{en,medicine}^{WN}$ refers to the set of English medical term candidates extracted from WN .

Compiling domain specific terminologies from open (domain free) resources, the objective of our work, is a challenging task clearly more difficult than obtaining them from domain restricted resources.

The adequacy of a resource for the task depends on the quantity of terms of the domain existing in the resource and on the difficulty of extracting them. Therefore, we can use as measures of such adequacy the following two metrics, whose values over several resources, for the domain of *medical drugs* for English, are shown in Table 2:

- Density of relevant terms in the resource, i.e. ratio between the number of tc occurring in the resource and the total number of terms. The higher this density is, the more adequate is the resource.
- Representing the whole resource as a graph (using the terms as nodes and the available relations between them as edges) and the occurring tcS as a subgraph, we can compute the ratio between the diameter of the subgraph and the diameter of the whole graph. The idea on using this metrics is that the more concentrated are the useful terms within the resource, the easier is to extract them.

Consider, for instance, on Table 2, the case of the rather specific domain of “medical drugs” for the English language. A terminology of English “medical drugs” can be obtained, from a highly specific resource as the DrugBank dataset⁷, that contains only drugs, [44], from a Medical domain resource as Snomed-ct⁸, that contains not only drugs but other medical terms as diseases, body parts and others, [38], or from open-domain resources as the English WN or the English DBP . Obviously, the resources closest to the domain of interest are the most adequate for extracting the terminology.

In Table 2, these measures are presented for the following resources for English:

- DrugBank

² Termhood refers to a degree of a linguistic unit to be terminological, i.e. the morpho-syntactic patterns valid for a term.

³ English: <https://en.wikipedia.org>, Spanish: <https://es.wikipedia.org>.

⁴ English: <http://wikidata.dbpedia.org/>, Spanish: <http://es.dbpedia.org/>. As a collective effort it is difficult to quote a single article for describing the content. [24] contains a comprehensive survey of WP uses.

⁵ English: <http://wordnet.princeton.edu/>, Spanish (within the Multilingual Central repository (MCR)): <http://adimen.si.edu.es/web/mcr/>.

⁶ The only linguistic processes in our system are tokenization and POS tagging.

⁷ <https://www.drugbank.ca/>.

⁸ <https://www.snomed.org/>.

Table 2 Comparison of resources for drug extraction. For all the datasets but DrugBank, we consider a full version, containing all the terms and a filtered version containing an identifiable fragment of the terms closest to the domain of interest

Resource	Size	Num drugs	Size ratio	Diameter	Diameter drugs	Diameter ratio
DrugBank	18,827	18,827	1.0	17	17	1
Full Snomed-ct	321,689	18,261	0.0567	35	30	0.86
Filtered Snomed-ct	4615	1312	0.2843	20	14	0.87
Full WN	82,115	461	0.0056	20	16	0.8
Filtered WN	914	293	0.32	8	5	0.63

- Full Snomed-ct
- Filtered Snomed-ct, i.e. terms in Snomed-ct under “pharmaceutical/biological product” top class.⁹
- Full WN^{en} (nominal WN)
- Filtered WN^{en} (nominal WN under drug synset: a substance that is used as a medicine or narcotic).¹⁰

Obviously, looking at Table 2, our first conclusion is that for getting the drugs in DrugBank the best is simply to download DrugBank. The two ratios have a value of 1 (the maximum). Using a medical resource as Snomed-ct could be a good solution in terms of coverage, most of the drugs in DrugBank exist also in Snomed-ct and the diameter ratio is not bad, but the size ratio is extremely low. It is worth noting that although the figures of the number of drugs are similar for DrugBank and Snomed-ct the intersection of both sets is small. The two resources are carefully curated but the criteria for inclusion and annotation are different. Brand names are included in DrugBank but not in Snomed-ct, many drugs are classified in Snomed-ct not as “pharmaceutical/biological product” but as “substance”. Filtering Snomed-ct and getting the terms occurring under the “pharmaceutical/biological product” top category gives a better result in terms of size ratio and the diameter ratio is also not bad but the coverage has dropped heavily, because many drugs are classified into other categories. The figures for WN follow a similar shape but at a level of coverage extremely low. The moral is that obtaining terminology for a domain using domain specific resources seems to be better than using generic resources. Unfortunately, there are no domain specific lexical resources for all our domains and, so, we have to follow an approach based on the use of domain-free lexical resources.

Our aim is using general (domain free) resources for a huge variety of domains, described in “[Normalising XWND info](#)”. We use as domain tags the set of WN domains, WND.

⁹ Snomed-ct is a bit tangled taxonomy with 18 top categories, between them “pharmaceutical/biological product”.

¹⁰ Also, the nominal part of WN is organized as a taxonomy with 13 tops. It is likely that descendant of a synset through hyponymy links consist of synsets of the same domain.

We restrict ourselves to extract terms corresponding to WP pages or categories. It is clear that depending on the language, the lexical source and the domain, the expected size of the tcS and the difficulty of the extraction task are different.

Our aim is collecting for each domain d in WND two sets of terms¹¹: English terms, $tS_{en,d}$, and Spanish terms, $tS_{es,d}$ as well as the possible mappings between them.

Obviously, not all English terms have Spanish counterparts (and also the reverse) and the mapping between the two languages is not always biunivocal. An English term can be mapped into more than one Spanish terms (and the reverse). If a bilingual mappings exist, we can take profit of such mappings for reinforcing the monolingual extractors, as discussed in “[DBP-based enrichment](#)”.

Our approach is extremely conservative in the sense of prioritising precision over recall. A set of thresholds have been used along the process for assuring an effective control of the confidence of the different components of the system. We set these thresholds in a very conservative way, although users are free to relax them for other uses of the system.

Some of the steps, namely 1, 4, 7, and 10 use thresholding mechanisms for allowing setting a balance between precision and recall. Although, as mentioned before, we have used an highly conservative setting in our implementation, other strategies could be used depending on the intended use of the terminologies.

We have used in our research English–Spanish mappings using as resources the inter-lingual links between English and Spanish WP (both between pages and categories). Other mappings are obtained from English and Spanish DBP, using “same_as” and “label” properties. Take into account our previous comments on the limited importance of the bilingual mappings in our system.

After this introduction, the organisation of the paper is as follows: “[State-of-the-art](#)” presents the state-of-the-art of four topics closely related to our system: WordNet domains, Extended WordNet domains, Ontologies, and Term Extraction. In “[General procedure](#)”, we describe in depth our

¹¹ Therefore, pages and categories whose title is not a term, as is the case of Named Entities, are discarded.

approach for extracting the terminologies of the domains. In “Applying the methodology: building the terminologies”, we present the terminologies obtained, their sizes and the representation formats. The evaluation framework is described in “Evaluation”. In “Results and discussion”, we present and discuss the results of the application of the evaluation framework to our data. Finally, in “Conclusions”, we state our conclusions and propose some lines of future work.

State-of-the-Art

We present in this section some topics closely related to our work. First, information sources, namely WN domains, extended WND and domain ontologies, next a brief survey of Term extraction techniques.

Information Sources

Dictionaries are collections of words in one or more specific languages, usually built for human use and usually arranged alphabetically, which may include information about definitions, usage, etymologies, pronunciations, translation, etc. In some cases, it includes also a field label for indicating a specific domain usage of such entry. This is the most common way to indicate in a systematic way the domain information for a given set of words. This information is added at the time of building the dictionary and usually it is not complete and systematical and rarely this information is updated to reflect new uses of a given word. Obviously, the completion degree closely depends of the type of dictionary: general or domain restricted.

As already mentioned in “Introduction”, domain terminologies (or dictionaries labelled with domain information) are a main resource required for many tasks in *NLP*. For this reason, there has been several attempts to obtain in a systematic way the terms of one or more domains. In the following subsections, we explore the more relevant projects in this area as well as some relevant resources that may help in (partially) facing this issue.

WordNet Domains

WordNet Domains, *WND* [23], is lexical resource where *WN* synsets (version 1.6) are annotated with domain information or Subject Field Codes (*SFC*) in words of the developers. *WN* nominal synsets have been annotated with *SFCs* by a semiautomatic procedure which exploits *WN* structure. The procedure started by annotating manually a small number of high level synsets. Then, an automatic procedure exploits some of the *WN* relations (i.e. hyponymy, troponymy, meronymy, antonymy and pertain-to) to extend the manual assignments to all the reachable synsets. The domain

hierarchy used to tag the synsets was specifically created to this purpose from a list of about 250 *SFCs* from a number of paper and machine readable dictionaries. Then, the list has been enriched on the base of the Dewey Decimal Classification, and later structured along two dimensions: inclusion, resulting in a *SFC* hierarchy, and semantic proximity, resulting in a number of *SFC* families.

Later, in [4], it was presented an updated version of this resource. It increases the number of *SFCs* to 164 and solve some issues mostly related to the lack of a clear semantics of the domain labels as well as the coverage and the balancing of the domains.

In [42], it was proposed a method for obtaining terms from potentially any domain using this resource together with Wikipedia. Such method chooses this resource as the reference domain taxonomy and the starting point to collect terms from a given domain. In this way, the resulting tool is language independent and could be applied to any language owning a relatively rich *WP*.

Extended WordNet Domains

Extended WordNet Domains¹², *XWND* (see [16]), is a resource, similar to that showed in the previous section, that uses a novel semi-automatic method to propagate domain information through the *WN*. It applies a graph-based method, based on the *UKB* algorithm¹³ [1] to generate new domain labels aligned to *WordNet* 3.0. *UKB* applies personalised PageRank on a graph derived from a wordnet.

In this way, it solves some issues derived from the updating process of *MCR* from *WN* 1.6 to 3.0. Such process leaves many unlabelled synsets (because there are new synsets, changes in the structure, etc.). Also, its use of the *UKB* algorithm for propagating information through a graph derived from *WN* structure allows to return a ranking of weights for each *WN* synset with respect to that particular domain. In this way, it is possible to know which is the highest weight for each domain and the highest weights for each synset. This allows to estimate which synsets are more representative of each domain (those who have more weight in the ranking) and which domains are best for each synset (those who have attained a higher weight for that synset).

Domain Ontologies

Ontologies are usually defined as an explicit specification of a conceptualisation (Gruber, 1993). They are artefacts allowing to represent explicitly the meaning of their components. Consequently, a domain ontology is a formal and

¹² <http://adimen.si.edu.es/web/XWND>.

¹³ <http://ixa2.si.edu.es/ukb/>.

consensual dictionary of categories and properties of entities of a domain and the relationships that hold between them.

Ontology building includes several steps. The most basic one is where the authors capture the knowledge of the domain and code it according to some formalism. The knowledge is represented using words; therefore, the building process can be seen as a simple collection of the words (or more precisely terms) relevant in the domain of interest. Later, such knowledge units are connected using a predefined set of relations.

There is a certain number of these resources covering one or more domains with different level of completeness. See below some examples:

- Biomedical domain: there is a large number of this type of resources in this area mainly because they are applied in the everyday tasks by health organisations. For example, the *BioPortal*¹⁴ provides access to more than 800 ontologies of this domain, including Snomed CT¹⁵, MeSH¹⁶, FMA¹⁷ and LOINC¹⁸ among many others.
- Law: the World Encyclopedia of Law.¹⁹
- Finance: The Financial Industry Business Ontology.^{20, 21}
- Multilingual/multidisciplinary thesaurus: the EU's EuroVoc²².
- Food and Agriculture: AGROVOC²³.
- Chemistry: CHEMINF²⁴.

Most of the above-mentioned resources are available only for English. This is a big issue in processing other languages. These examples can be useful as lexical sources for building terminologies for the corresponding domains, probably leading to better results, as discussed in “Domain ontologies” but, unfortunately, we are far from disposing of appropriate domain resources for covering the whole set of 164 WND domains. This is why we moved to general domain-free lexical resources for our purposes.

Term Extraction

Usually, terms are defined as lexical units that designate concepts of a thematically restricted domain. Obtaining the terminology of a domain may be a necessary requirement for a large number of tasks such as building/enriching/updating lexical repositories, ontology learning, summarisation, named entity recognition or information retrieval among others. At the same time, it is problematic task. Two problems arise: first, a well organised corpus of texts representative of the domain should be available and, second, extracting the terms of such corpus.

Compiling a corpus is an expensive task in time and resources. Obtaining the terms of the corpus already compiled is also problematic. Manual processing is an unattainable task, and thus automatic methods are used, although, the results are not perfect. Therefore, term recognition constitutes a serious bottleneck.

Since the nineties this task has been object of research but in spite of the efforts, it cannot be considered a solved issue. We can see the extraction of the terms as a semantic labelling task by adding 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 would not be available, it is necessary to resort to indirect information of a linguistic and/or statistical nature. The results obtained with these mechanisms 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 semantic resources (see [7] and [32]), many indirect methods have been proposed to obtain the terms included in texts. Some of them are based on linguistic knowledge, like in [17], others use statistical measures, such as ANA [10]. Some approaches combine both linguistic knowledge and statistics, such as TermoStat [9] or [14]. Linguistic methods are based in the analysis of morpho-syntactic patterns but in this way the result is noisy and terms candidates cannot be scored, leaving the relevance measure to the domain experts. Statistical methods are more focused in detecting general terms relying in the frequency as its basic measure. Terms candidates may be ranked but again an expert is necessary to measure its actual relevance.

Recently, [22] propose a full 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 extract terms integrating, by design, both term extraction and term classification tasks. These methods use to require training data. The lack of reliable tagged resources for training

¹⁴ <https://biportal.bioontology.org/>.

¹⁵ Systematized Nomenclature of Medicine Clinical Terms, <https://www.snomed.org/>.

¹⁶ Medical Subject Headings <https://www.nlm.nih.gov/mesh/>.

¹⁷ Foundational Model of Anatomy <http://sig.biostr.washington.edu/projects/fm/AboutFM.html>.

¹⁸ Logical Observation Identifiers Names and Codes <http://loinc.org/>.

¹⁹ <https://lawin.org/estrella-project/>.

²⁰ <https://edmcouncil.org/>

²¹ FIBO, <https://edmcouncil.org/page/aboutfiboreview>.

²² <http://eurovoc.europa.eu/>

²³ <http://aims.fao.org/vest-registry/vocabularies/agrovoc>.

²⁴ Chemical Information Ontology, <http://aber-owl.net/ontology/CHEMINF/>.

Fig. 1 The Wikipedia graph structure

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 [8] and [27]. Also, recent techniques like those using deep learning have been applied to the task; see [43].

As already mentioned, a common limitation of most extractors is the lack of semantic knowledge. Notable exceptions for the medical domain are MetaMap [2] and YATE [7]. Most approaches focus on technical domains in which specific resources are available²⁵ and term recognition is a bit easier. As medical 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 use to be syntactically ill formed.

Recently, deep learning models have achieved great success in fields as computer vision and pattern recognition among others. *NLP* research have also followed this trend as shown in [46]. This success is based on deep hierarchical features construction and capturing long range dependencies in data. These techniques have also been applied in electronic health reports for clinical informatics tasks (see [37] 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) is an example of such approaches in the Biomedical domain using approximate matching for increasing the coverage of the system.

General Procedure

From the three lexical sources used in our approach, the basic one is the *WP*, because our aim is extracting terms occurring as titles of *WP* pages or categories, while the other two are somehow complementary; *WN* provides the domain taxonomy and the initial seeds for *tcS* and *DBP* allows both monolingual and multilingual enrichment of the *tcS*.

We obtain the terminology for a domain using for each language *l* two *WP* graphs as knowledge sources, the graph of *WP* pages, WP_{PG}^l , and the graph of *WP* categories, WP_{CG}^l . Figure 1 shows these two graphs and their relations.

Our hypothesis is that page and category titles are good *tc*. From WP_{PG}^l and WP_{CG}^l graphs, we use the following types

Fig. 2 Full pipeline

²⁵ Note that most of the available resources refers to English.

Fig. 3 Obtaining WP categories

Fig. 4 Extracting WP pages

Fig. 5 Obtaining and refining final terms

of edges: page $\rightarrow$ category and the inverse, category $\rightarrow$ category (super and sub-categories), and its inverse, page $\rightarrow$ page (input and output links from/to a page text).

We use as seed categories for starting the process the variants included in the synsets of the corresponding WN likely belonging to the domain, as explained in “Getting WN and MCR variants”.

Roughly, as shown in Fig. 2 the system proceeds into three steps:

1. Selection of valid categories
2. Selection of valid pages
3. Selection of valid terms

The overall process, outlined in Fig. 2 and detailed in Figs. 3, 4, and 5, is iteratively applied to each pair $\langle d, l \rangle$ independently. Let the pair $\langle dc, l \rangle$ being $dc \in WND$, i.e. the dc assigned to d . The process consists on the steps listed below that are described shallowly on the following sub-sections and in depth in “Detailed description of the pipeline”.

Selection of Valid Categories

In this module, we select for each domain and language the set of relevant categories of the corresponding WP^l . These categories could be useful as term candidates and as collections of relevant pages in WP^l . This module consists of the following six steps (see Fig. 3):

- Step 1) (see “Normalising XWND info”) We choose the semantic tagset, i.e. the set of domains for which terminologies have to be extracted. As mentioned before, we decided to use $XWND$ after doing some data normalisation.
- Step 2) (see “Getting WN and MCR variants”) For each dc and l , the process starts by locating, using MCR together with $XWND$, all the synsets in WN^l having a high probability of belonging to such dc . Then, we look for the WP categories associated to each variant. Such

list of *WP* categories are evaluated and those having a probability to belong to the domain lower than a given threshold are filtered out.

- Step 3) (see “[Getting WP Top Categories for Each *dc*](#)”) In this step, we obtain the *WP* top categories corresponding to each *dc*.
- Step 4) Using a threshold, we obtain initial sets of in-domain ($catS_i^{l,dc,init+}$) and off-domain ($catS_i^{l,dc,init-}$) categories. The process is explained in “[Collecting Initial set of WP categories](#)”.
- Step 5) (“[Top-down expansion of WP categories](#)”) The categories in the initial set of candidate categories, $catS_0^{l,dc,init+}$, are top down expanded traversing WP_{CG}^l following the subcategory links for obtaining the set of expanded categories, $catS_0^{l,dc,expanded+}$.
- Step 6) Applying a set of selecting methods, *m*, described in “[Generating initial valid categories set](#)”, we obtain an initial set of valid categories for each language *l*, each domain *dc* and each selecting method *m*.

Selection of Valid Pages

From the selected categories, this module tries to extract the set of relevant pages. It consists of the following three steps (see Fig. 4):

- Step 7) “[Iterative category and pages selection](#)” explains how the initial set of pages, $pageS_i^{l,dc,m}$, is built. From each category in the set of initial candidate categories, $catS_0^{l,dc,m}$, the belonging pages, following *cat* → *page* links, is collected.
- Step 8) An iterative process is carried out²⁶. In iteration *i* the sets of categories, $catS_{i-1}^{l,dc,m}$, and pages, $pageS_{i-1}^{l,dc,m}$, of the previous iteration constraint or reinforce each other for getting the new sets of categories and pages, $catS_i^{l,dc,m}$ and $pageS_i^{l,dc,m}$. The process is detailed in “[Iterative categories and pages term generation](#)”.
- Step 9) The best combination of *m* and *i* is chosen as shown in “[Final candidate term selection](#)”. In this way, the initial set of term candidates from the *WP*, tcS_l, d^{wp} is built.

Selection of Valid Terms

Once the sets of valid categories and valid pages are built, we face the final problem of selecting from these sets the final terminologies for each language and domain. The process consists of the final three steps (see Fig. 5):

Table 3 Normalised weight rankings for the noun sense < drug₁ >

Normalised weight	Domain
68.62	PHARMACY
11.55	FREE-TIME
7.83	COMMERCE
3.39	PHYSIOLOGY
3.34	CHEMISTRY

- Step 10) A pagerank-based algorithm is then applied over a graph representation of the selected *tcS* for scoring the candidates and filtering out the less reliable. See “[PageRank-based refining](#)”.
- Step 11) Till now, all the steps have been applied independently for each <*dc*, *l*> pair. There are, however, problems of overlapping, the most serious one the case of *tc* belonging to more than one incompatible *dc*, that have to be addressed. We face these problems in “[Dealing with global issues](#)”.
- Step 12) Finally, the set of terms can be enriched using *DBP^l* for both languages. Two kind of enrichment are considered, monolingual and bilingual. The process is explained in “[DBP-based enrichment](#)”.

Some of the steps, namely 1, 4, 7 and 10 (see Figs. 3, 4 and 5), use a thresholding mechanism for allowing a control over the balance between precision and recall. For building our terminology, we have followed a highly conservative approach but other strategies could be followed depending of the intended use of the resource.

Detailed Description of the Pipeline

The following sections provide a more detailed description of the modules already mentioned.

Normalising XWND info

We use as domains those defined in *WND/XWND*²⁷ because this resource provides a complete and well known domain hierarchy. Also, it is possible to obtain domain indicators for each synset/variants included in *MCR*. We use this resource for obtaining the seed terms for each domain and later as a cheap, though partial, evaluation resource of our results.

As already mentioned in “[State-of-the-art](#)”, *XWND* includes a weight to quantify the pertinence for every synset to all domains defined in *XWND*. However, such weights should be normalised across domains to be useful in our task. A normalisation procedure has been done in two steps.

²⁶ The process iterates until no changes on the sets are obtained. In none of our domains, the number of iterations have been greater than 7.

²⁷ We have excluded the *dc* ‘factotum’

Table 3 shows the first five domains and their normalised weights resulting from the application of this method to synset < drug₁ >. This synset, the first sense of “drug”, is defined as “a substance that is used as a medicine or narcotic” and is originally labelled in *WND* just as PHARMACY. The algorithm used for building *XWND* improve the initial labelling, because it suggests PHARMACY (the best option) like *WND* but also other related senses such as PHYSIOLOGY and CHEMISTRY among others. Probably, the domains FREE-TIME and COMMERCE collect the sense corresponding to narcotic.

Getting WN and MCR Variants

In practice, this procedure consists of finding the domain name in EWN and consider it as the domain border. Then, we collect all the variants that exist under this border. Our approach started with English, because *XWND* exists for this language. Using the interlingual index (*ILI*) existing in *MCR*, we are able to map these synsets into to the corresponding Spanish ones. The process follows from now on in parallel for both languages.

The next step is to process the whole set of variants with wikiYate (a term detection system, see [41] and [6])) to evaluate its termhood. The purpose of this step is to keep only those strings that are considered valid terms in the domain under consideration and to obtain its *WP* categories. Such list of categories, due to the characteristics of *WP* may contain out of domain categories. To avoid invalid expansions such list of categories is analysed with wikiYate. Only those which termhood value is above a threshold are accepted. The

As an example consider the domain *agriculture*; its WP page assign three categories: *Agriculture*, *Agronomy* and *Food industry*. The category *Food industry* will be discarded due to its low termhood value in the field of agronomy. The same situation may arise in others domains and languages. If we consider the Spanish term *arquitectura* (architecture), we would see that its WP page assign to this page three categories: *Arquitectura*, *Arte* (Art) and *Construcción* (Construction). As in the previous example there is a WP category that will be discarded (architecture) due to its low termhood in this domain. In both cases, the categories that remain allow the terms of each domain to be properly collected.

In this step, we obtain the top categories in WP_{CG}^I corresponding to dc . We get this set using with decreasing confidence WP_{CG}^I , WP_{PG}^I , page-category edges and interwiki edges. In most of the cases, e.g. “Medicine”, dc directly corresponds to a category in WP_{CG}^I , so top categories for this dc consisted on just {Medicine}, in other cases, however, more than one top is needed. For instance, for English the dc for the tops of the domain “Economy” include the categories {Economic_systems, Economics, and Economies}.

Now, we extract from WN^l , all the variants contained in all the synsets tagged in normalised $XWND$ with dc . We analyse these variants, obtaining from each their set of categories,

²⁸ As a matter of fact, our thresholding mechanism takes into account the normalised weight of the first domain associated to the synset and the difference with the second one.

Fig. 7 The WP categories related with the WP page “Asthma”

using the page $\rightarrow$ category links, resulting on the initial set of candidate categories. Using a threshold, we obtain an initial set of in domain categories and a complementary set of off-domain categories. For “Medicine”, the sizes of these sets were respectively 253 and 2263.

Top-Down Expansion of WP Categories

For a term candidate corresponding to the title of a WP category, there are two ways for looking for new terms: (i) following the cat $\rightarrow$ page links, i.e. looking for terms associated to WP pages, and (ii) following the sub-category links. Both types of links use to be quite robust. We follow in this step the later approach. If a WP category denotes a set of WP pages, a subcategory denotes one of its subsets.

Now, categories in the initial set of candidate categories are top down expanded traversing WP_{CG}^I following the subcategory links, avoiding cycles, filtering out neutral categories²⁹ and categories placed in WP_{CG}^I above the domain tops and discarding the expanded categories and descendants when belonging to the set of the initial candidate categories.

In this way, the final set of expanded candidate categories is obtained. For “Medicine”, the size of this set was of 1,924 categories.

²⁹ In WP, some categories are used for monitoring/grouping the state of some pages not yet complete (e.g. “All articles lacking sources”, “Articles to be split”, “Hidden categories”, “Commons category link is on Wikidata”, ...). We name such categories as *Neutral Categories* and are not used for expansion

Generating Initial Valid Categories Set

We get in this step from the set of expanded candidate categories the first set of candidate categories. For building this set, we apply some filtering processes over the candidates taken into account the whole set of category candidates (turquoise shaped zone in Fig. 6) the set of categories under the top categories of the domain (black shaped zone in the figure) and the paths from the category candidate c to these tops, noted p_1 , p_2 , p_3 , and p_4 in the figure.

Using these Knowledge Sources, we can filter out (i) candidates unconnected to the top categories of the domain (likely not included into the domain), (ii) candidates placed higher than tops in WP (likely to be too general terms), (iii) low scored candidates in the case of huge size of the expanded candidates set, and (iv) low scored candidates in the case of low domain ratio.

Consider as an example the medical domain shown in Fig. 7. Some of the most relevant WP categories associated to the page “Asthma” are shown. The left part of the figure depicts the “Asthma” page at the bottom and other WP pages above. The “Asthma” page is directly connected with 5 WP categories (block 1). In the middle of the figure, the top category of the domain, “Medicine” occurs. The categories in block 1 are connected to the top of the domain through several paths presented in block 2. The categories in block 2 are far to follow a taxonomic structure; cycles and links to categories outside the scope of the top are frequent. The absolute top of the $WP_{categories}^{en}$, “Main topic classifications”, is shown at the top of the figure. Block 3 presents the best

Table 4 Topologic al properties of categories in $WP^{en}_{categories}$

WP category	Domain coverage	Global coverage	Domain ratio	Distance to domain top	Distance to global top
Main topic classifications	59,874	890,911	0.067	– 1	0
Science	13,020	162,755	0.08	– 1	2
Medicine	59,874	59,874	1	0	3
Life sciences	20,955	59,874	0.35	– 1	5
Health sciences	100,908	162,755	0.62	– 1	6
Natural sciences	19,291	120,572	0.16	– 1	4
Asthma	8	13	0.64	4	10
Otorhinolaryngology	1043	2981	0.35	3	5
Respiratory therapy	21	31	0.69	4	6
Clinical medicine	15,418	22,026	0.7	1	4
Autoimmune diseases	36	44	0.82	7	9
Respiratory system	728	2208	0.33	– 1	11
Pulmonology	257	330	0.78	3	5

path between the absolute top and the top of the domain. The topological properties of some of these categories are presented in Table 4. For each category c in column 1, we represent the size of the sub-graph headed by c in column 3, the size of its sub-graph under the topic top category in column 2, the ratio of these two sizes in column 4, the distance from c to the topic top category in column 5 (– 1 if no connection exists) and the distance to the global top in column 6. For computing such distances, we have used the shortest paths between the categories using the subcategory links.

Using the criteria defined above, we can filter out all the candidates unconnected to the domain top (those having a – 1 in column 5 of the table. We discard too, the candidates whose distance to the global top is lower than the distance from the domain top to the global one (3 in our case). We discard too the candidates with a domain ratio smaller than a threshold (set to 0.5 in our experiments). Finally, in the case of very huge candidate sets, we maintain the candidate set but removing from it the less scored categories.

Iterative Category and Pages Selection

The initial set of pages is built (for all the methods m considered above). From each category and for both languages, the set of belonging pages, following $cat \rightarrow page$ WP links, is collected. The $cat \rightarrow page$ links are in general many to many, so a page candidate could belong to more than one category. We define a purity score of a page as the ratio among its valid categories and all the categories attached to the page. We restrict the obtained pages to those whose purity is over a purity threshold. In our setting, we use a purity threshold of 1, i.e. we discard all the pages but those absolutely pure but other settings can be used instead. Consider again the example of Fig. 7, the page “Asthma” is connected to 5 categories,

all of them valid, so the purity of the page is 1 and the page is selected as valid. The category “Asthma” contains, however, 36 other pages, as “Brittle asthma” or “World Asthma Day”. The first one has also a purity 1 but this is not the case for the other. “World Asthma Day” belongs to the category “May observances”, clearly out of domain. Each category is scored according to the scores of the pages it contains and each page is scored according to both the set of categories it belongs to and to the sets of pages linked to it. Three thresholding mechanisms are used:

- *Microstrict* (accept a category if the number of member pages with positive score is greater than those with negative score),
- *Microloose* (similarly with greater or equal test), and
- *Macro* (instead of using the page scores, we use the scores of the categories of the pages).

The general scoring mechanism combines the purity threshold and the filtering method m . In our setting from the different combinations, we use the one most restrictive, i.e. the one accepting the minimum number of terms. We obtain in this way the set $pageS_0^{l,d,m}$

Iterative Categories and Pages Term Generation

Then, in step 8, we iteratively explore each category repeating the same process again. At iteration i , the set of well scored pages and the set of well-scored categories reinforce each other. Less scored categories and pages are removed at each iteration, so the global precision of the sets is expected to grow at a cost of a draw in recall. A combination function is used for computing the global score of each page and category from their constituent scores. The process is

iterated until convergence. These sets are collected for all the iterations and selection methods.

Final Candidate Term Selection

In step 9, a final filtering is performed for selecting from all the categories and pages corresponding to all the iterations and selection methods in step 6 the one with best F1.

PageRank-Based Refining

The aim of this step is a further refining of the dataset of terms taking into account their interrelations. To this end, we use a centrality measuring approach. We start representing the current dataset, i.e. the set of pages and categories selected in step 9 as an un-directed graph including all the terms as nodes and including as edges the relations between the terms attached to *WP* pages and categories. The process of filtering is carried out over the candidates for the two languages independently. We examined several centrality measures, as degree, closeness, and betweenness, [28], testing them on the set extracted for the medical domain. The best measure resulted the PageRank algorithm [31], so in this step, we apply it to all the domains and languages. We represent the set of terms as nodes of the graph and add iteratively new edges in the following way³⁰:

1. Including edges between nodes representing pages and categories using the *cat* → *page* and *page* → *cat* links.
2. Including edges for input and output links occurring within the text of *WP* pages linking pages already included within the graph.
3. For nodes corresponding to multiword terms, the set of simpler constituents is obtained and for those existing as nodes in the graph a new edge linking the complex and simpler nodes is built.
4. For nodes of type *cat*, links to sub-category, super-category and distance 2 links are included, just in the case the linked categories exist in the graph.

The incorporated edges are weighted linearly according to the step triggered before in descending order. Therefore, edges added by 1) are weighted with 1., those added by 2) with 0.8, and so on. Nodes are then scored using a non-directional variant of the PageRank algorithm, first proposed in the *TextRank* algorithm [25] and less scored nodes are removed from the dataset. To set a threshold for selecting the final set of terms, we considered the domain of “Medicine” for which golden datasets are available (we used *Snomed-ct*

for English and Spanish). We built an histogram showing the coverage of terms included in *Snomed-ct* for different values of the threshold and we set the threshold to 80%. Then, we manually validated this threshold using a sample of 1000 terms for each language in the domain of “Medicine” confirming the adequacy of the threshold. This setting is, however, rather brittle, and a more in depth analysis of the adequacy of a threshold, including the assignment of a different threshold for each domain should be performed. This analysis needs the availability of golden datasets, at least for some domains and is left as future work.

Dealing with Global Issues

Previous steps have been applied independently for each (*dc*, *l*) pair, and, to some extent, terms corresponding to pages and to categories have been extracted independently. In this step, all the terms collected for each language are analysed together to detect overlapping between *dc* and inconsistencies. The following issues are analysed and corrected:

- Term duplication for a category and a page. When the same word form occurs the one corresponding to the category is removed.
- Term duplication with different wording: different case or hyphenation, addition of parenthesised tags, etc. Terms are normalised and only the canonical one is kept. We follow for this normalisation the *WP* guidelines for titles, terms are capitalised and components of multi-word terms are separated by spaces.
- Existence of terms in singular and plural form³¹. If both exist, the singular form is kept. Plural form is only maintained in the case of a category whose page form is singular.
- The most serious issue is the case of *tc* belonging to more than one incompatible *dc*. In this case, we look for compatibility between the *dc* associated to the *tc*. For addressing this case, we have clustered *dc* in *WND* into compatible clusters. Although ideally *WND* is a true taxonomy, i.e. two *dc* with no common ancestors are assumed to be incompatible, this is not always the case and some *dc* allow some degree of overlapping. For instance, “pharmacy”, “biology”, and “anatomy” are clustered together with “medicine” as a compatible set of tags. The taxonomic structure of *WND* has been used for the initial assignments that was completed manually. In the case, a term belongs to more than one *dc*, the most likely cluster, i.e. the one containing more *dc* associated

³⁰ Note that the process simply enrich the graph with new edges, the set of nodes remains unchanged.

³¹ Number detection has been performed using Python *Inspector* module, “<https://pypi.org/project/python-inspector/>”, that provides tools for English and Spanish.

to the tc , is chosen and all the dc not belonging to the cluster are removed³². Consider, for instance the case of the term “Antibiotic” that only belongs to “Medicine” and “Pharmacy”. As these two domains belong to the same compatible set, both assignments are kept.

- Sometimes, some of the dc associated to the tc are so frequent that likely correspond to pathological (i.e. no terminological) assignments. This is the case of “person” for Spanish or “economy” for English. In this situation, the assignment to such dc are simply removed.
- A last case occurs when a tc has many dc associated to it³³. This case could correspond to highly ambiguous or very general terms. In this case, the term is removed from all its assignments.

In this way, the set of tc is cleaned from probably erroneous terms.

DBP-Based Enrichment

Finally, we try to use both versions of DBP^l for enriching the $tS_{l,d}$.

We use a similar mechanisms for monolingual (new terms for the same language) and bilingual (terms in the other language) enrichment. Two mechanisms are used, the former is based on the property “label” that assigns to a DBP^l resource a set of words that can be considered synonyms of the resource in several languages. The latter is based on the property “same_as” that links resources in several datasets. We use this property for linking DBP^{en} and DBP^{es} . We use this mechanisms in a very restricted way, constraining the obtained terms to correspond to pages or categories in WP^l . The new terms are added to the collection for finally obtaining tS_l , for $l \in \{en, es\}$.

Applying the Methodology: Building the Terminologies

Applying the methodology explained, and using highly conservative thresholds, for focusing on accuracy, we were able to collect a total of 635,527 terms covering the 164 domains for both languages³⁴. In Table 5, we present the ten domains owning the most terms. The last row shows the overall figures. The table shows for each domain the number of terms corresponding to pages and categories for English and Spanish and the total number of terms. We include as

³² Frequency of the dc is considered in the case of ties.

³³ More than 5 in our experiments.

³⁴ These datasets will be made public in the case of acceptance of the article.

Table 5 Most frequent dc extracted terms

dc	en pages	en cats	es pages	es cats	mappings	total
social	41,710	2230	6206	808	6583	50,954
free_time	26,819	461	1223	136	716	28,639
animals	16,281	636	6936	206	4661	24,059
person	17,163	589	5502	293	3100	23,547
biology	13,847	754	4318	339	3036	19,258
medicine	13,353	852	4227	423	3473	18,855
plants	5366	271	10,436	1428	472	17,501
environment	14,124	901	2105	235	1996	17,365
sociology	13,715	1315	1874	452	1738	17,356
industry	13,774	229	2165	215	1020	16,383
.						
total	444,653	28,032	146,140	16,702	79,946	635,527

well (the mapping column) the number of terms for which a translation exists. The results, compared with sizes of the 2 involved *WP*: 5,809,613 English pages, and 1,506,247 Spanish pages³⁵ seem to be low at first glance but we have to take into account the encyclopedic character of the *WP* that results on a rich representation of Named Entities that cannot be considered as terms, according to the termhood constraint.

We selected a set of six domains for performing the evaluation.³⁶ We include in this set the domain “Medicine” for the sake of relative comparability with Table 2. In Table 6, we present some metrics of the graph representations of the terms produced by our system for these evaluation domains, for English and Spanish. This graph representation was produced in an intermediate stage of our system as described in “PageRank-based refining”.

Some interesting results can be obtained from Table 6. There are two pathological cases: “Music” for English and “Tourism” for Spanish, more than one order of magnitude below the other domains. For the other domains, the coverage (number of nodes) is good. The case of “Medicine” for English is relatively comparable with the figures of Full Snomed-ct in Table 2. Our results, 14,205 terms, are obviously lower than the 321,689 terms occurring in Snomed-ct. This result was expected, because Snomed-ct is a specific resource for the medical domain and it is not the case for our sources. Besides, the diameter is better and most of the Snomed-ct terms do not occur in *WP*. Comparison between diameter ratios have no sense because the type of edges in our case and in the whole *WP* are different.³⁷ For all the domains, the diameter of the graphs are around ten edges, clearly below the values depicted in Table 2.

The whole terms collection has been saved on a set of 164 files following OLIF³⁸ formatting norms. OLIF, the Open Lexicon Interchange Format, is an open standard for lexical/terminological data encoding. Figure 8 shows a fragment of such files. The figure shows the description of the concept “med_4434”, from the medical domain. which is realised both in English, as *assisted reproductive technology* and its Spanish translation, as *reproducción asistida*. The two terms are mapped to each other therefore they have the same identifier. The entry descriptor includes some complementary information; the POS, the source, *WP* page or category, and

³⁵ Figures obtained from https://en.wikipedia.org/wiki/List_of_Wikipedias.

³⁶ The reason for choosing these domains is presented in “WordNet domains”

³⁷ In a curious article, <http://mu.netsoe.ie/wiki/>, Stephen Dolan analyses the problem of computing the diameter of the *WP*. He detects “giant” tails, up to 70 long chains of almost linear linked lists of pages, that invalidate the usual definition of diameter.

³⁸ <http://www.olif.net/>.

Table 6 Some metrics of the graph representations of the terms produced by our system for the evaluation domains

domain	l	Density	Diameter	Num. nodes	Num. edges
Agriculture	en	13.97	9	8813	123,084
Anthropology	en	14.24	11	2630	37,454
Architecture	en	6.88	10	8271	56,900
Medicine	en	10.13	10	14,205	143,826
Music	en	1.97	7	573	1126
Tourism	en	3.73	10	3517	13,130
Agriculture	es	2.50	12	901	2254
Anthropology	es	15.47	9	1400	21,656
Architecture	es	3.40	12	1961	6664
Medicine	es	13.43	8	4650	62,450
Music	es	12.40	9	1393	17,262
Tourism	es	0.47	7	537	250

Fig. 8 Example of a term and its Spanish translation in the OLIF format

<pre> <entry ConceptUserId="med_4434"> <mono> <keyDC> <canForm>assisted reproductive technology</canForm> <language>en</language> <ptOfSpeech>noun</ptOfSpeech> </keyDC> </mono> <monoDC> <monoAdmin> <confidence>0.801</confidence> <entrySource>WP page</entrySource> </monoAdmin> </monoDC> </entry> </pre>	<pre> <entry ConceptUserId="med_4434"> <mono> <keyDC> <canForm>reproducción asistida</canForm> <language>es</language> <ptOfSpeech>noun</ptOfSpeech> </keyDC> </mono> <monoDC> <monoAdmin> <confidence>0.800</confidence> <entrySource>WP page</entrySource> </monoAdmin> </monoDC> </entry> </pre>
--	---

the confidence score. The full collection is publicly available for download.

Evaluation

In this section, we present our evaluation framework. The results of applying this framework to our data are presented and discussed in “[Results and discussion](#)”. Generally speaking, evaluating a terminology is a difficult task due to the:

- difficulty in doing it through human experts (therefore it becomes a subjective task),
- lack/incompleteness of electronic reference resources and
- disagreement among them (specialists and/or reference resources).

As said above, we decided for the evaluation selecting six domains. We sorted the set of domains by the count of terms selected and we splitted the list into 3 bulks, less than 3000 terms, between 3001 and 8000 and more than 8000. We removed from these lists the domains classified in different bulk for the two languages and we selected randomly³⁹ a pair of domains from each bulk: “Anthropology” and “Music” from the first bulk, “Architecture” and “Tourism” from the second bulk, and “Agriculture” and “Medicine” from the third.

Looking to minimise the above mentioned problems, we set up four different scenarios for evaluating the resources obtained with our system. All the evaluations have been carried out on the set of six domains chosen from the task as indicated in “[Domain ontologies](#)”. The following are the evaluation scenarios designed for evaluating such domains:

- Due to the lack of reliable references for most of the domains, we planned a first scenario doing a partial evaluation, restricted to terms occurring both in *WN* and *WP*, using as golden dataset the assignments included in *WND*. This evaluation setting can be applied to any domain/language.

Fig. 9 Terms indirect evaluation

³⁹ The medical domain, for which some material for evaluation was available, at least for English, was directly included in the selection.

Table 7 Global datasets sizes for Magnini and NG baselines

dc	WP	WN	WP $\cap$ WN	WNall
Agriculture	8640	1495	1392	1434
Anthropology	2258	1109	1055	1174
Architecture	7850	5542	5480	5677
Medicine	13,242	11,795	11,088	11,143
Music	206	2002	1852	1900
Tourism	3437	1494	1400	1509
all 6 dc	35,633	23,437	22,267	22,837

- Instead, for those few domains where an external reliable reference is available a full evaluation scenario was foreseen. This is the case of “Medicine” for which we use *Snomed-ct* [38].
- Limited to the English language and to terms included in *WP* we perform a comparison with a hard baseline system: Niemann and Gurevych, *NG*, [29].
- The last scenario is an indirect one. We use the content of the *WP* pages associated to the terms of a domain for learning word embeddings for such domain for further evaluating the embeddings.

The following sections describes such scenarios with some detail.

Basic Evaluation Scenarios

Our guess is that accuracy results can be extrapolated to terms not occurring in *WN* and to domains lacking external references. First, consider Fig. 9 that shows the basic sets of *TCs* that must be considered in doing this evaluation. This evaluation is based on the subset *C* as they are the only terms that are tagged in *WN* as belonging to the domain and at the same time should take part of the set of pages/categories found in *WP*. Precision and recall can be easily calculated as indicated in formulae 1 and 2.

$$\text{Precision} = |C| / (|C| + |D|), \quad (1)$$

$$\text{Recall} = |C| / (|B| + |C|) \quad (2)$$

$$F1 = 2 \times (\text{precision} \times \text{recall} / (\text{precision} + \text{recall})). \quad (3)$$

The four evaluation scenarios are detailed in next subsections.

Comparison with *WND* and *NG* Baseline Systems

An additional evaluation has been performed, for English, comparing our results with two hard baselines, one based on *WND* (Magnini [23]) and the other based on the alignment

of *WN* senses to *WP* pages (Niemann and Gurevych, *NG*, [29]), described next.

- *WNdomains*. This baseline consists simply on, giving a domain code, *dc*, of Magnini’s taxonomy, collecting all the synsets of *WN* assigned to it. Then, we consider as terms of the domain all the variants associated to these synsets whose normalised weight (according *XWND* info, see 3.4.1) is higher than a threshold. We experimentally set this threshold to 0.1. This approach has the obvious limitation of reducing coverage to the variants contained in *WN*; also it is rather crude, because no score is attached to *TCs*, despite their degree of polisemy or domainhood. The dataset we have used is the version *wn-domains-2.0-20050210*⁴⁰ that maps synsets of *WN1.6* to one or more *dc*, in Magnini’s taxonomy. The dataset been further mapped from *WN1.6* into *WN3.0* using the *WN* mapping tools provided by TALP.⁴¹
- *NG*. This system maps *WP* pages with *WN* synsets reaching a 0.78 F1 score. Our baseline is built collecting all the synsets corresponding to *dc* and from them all the *WP* pages aligned with the synset. The dataset consists of 274,870 mappings from *WP* pages into *WN3.0* synsets.⁴² Two scores are attached to each mapping one measuring word overlapping and the other based on *Personalised PageRank*, [1]. Mappings are tagged as ‘accepted’ when a threshold of 0.048 for the word overlap score and 0.439 for the *Personalised PageRank* score are satisfied. Look at [29] for details about these two thresholds and their use.

We have collected the set of terms selected by *NG* baseline for comparing this set with the set of terms extracted by our method and with Magnini’ dataset. For the sake of comparability, the terms are normalised into a common format. The normalisation includes lowercasing, using space as delimiter in the case of multiword terms and removing of parenthesised categories occurring sometimes as suffixes of the terms in *WP* pages titles. The collections correspond to the six *dc* analysed before (‘Agriculture’, ‘Anthropology’, ‘Architecture’, ‘Medicine’, ‘Music’, and ‘Tourism’). The evaluation has been performed for English.

For each *dc*, the set of terms occurring in *WP*, *WN* and their intersection are collected. For providing an insight of the coverage of *WN* by *WP*, a column *WNall* showing the

⁴⁰ <https://github.com/larsmans/wordnet-domains-sentiwords/blob/master/wn-domains/wn-domains-2.0-20050210>.

⁴¹ <http://www.talp.upc.edu/content/wordnet-mappings-automatically-generated-mappings-among-wordnet-versions>.

⁴² https://www.informatik.tu-darmstadt.de/ukp/research_6/data/index.en.jsp.

Table 8 Sizes of the datasets used for the baseline evaluation scenario

dc	WP	WN	$WP \cap WN$	WP	WN	$WP \cap WN$
	Agriculture			Anthropology		
ALL	9129	1495	1392	3147	1109	1055
MAGNINI	504	383	280	906	415	361
NG0	268	239	239	224	209	209
NG4	268	239	239	224	209	209
NG5	210	189	189	200	187	187
NG6	50	48	48	62	61	61
OURS	8640	1733	1733	2258	345	345
	Architecture			Medicine		
ALL	12,978	5542	5480	22,867	11,795	11,088
MAGNINI	5251	209	147	10,763	3392	2685
NG0	1178	1144	1144	2500	2405	2405
NG4	1178	1144	1144	2500	2405	2405
NG5	1014	984	984	2220	2137	2137
NG6	317	311	311	748	730	730
OURS	7850	769	769	13,242	2915	2915
	Music			Tourism		
ALL	2003	2002	1852	4734	1494	1400
MAGNINI	1817	1217	1067	1309	557	463
NG0	356	354	354	299	291	291
NG4	356	354	354	299	291	291
NG5	322	320	320	250	244	244
NG6	135	134	134	82	81	81
OURS	206	117	117	3,437	219	219

count of *WP* terms occurring as variants of *WN* synsets, regardless their *dc* assignment, is included as well. The figures regarding the sizes of these sets are shown in Table 7. The global datasets are built as follows:

- The *WN* collection includes the terms included in Magnini dataset. For getting these terms from the synsets, we used the *WN* module of NLTK.⁴³
- The *WP* collection includes all the terms proposed by the three systems, namely:
 - The set of terms in the *WN* collection existing in *WP* (as pages or categories). We have used python *wikitoools* for this task.⁴⁴
 - The set of terms included in NG dataset
 - The set of terms extracted by our system.
- $WP \cap WN$ is simply the intersection of *WP* and *WN* collections. As most of the *WN* terms exist in the *WP*, this set is almost identical to *WN* dataset.
- *WN*all contains the intersection of the *WP* dataset and the whole nominal *WN*. As Magnini's has a rich coverage, this set has a size only a bit higher than $WP \cap WN$.

Table 9 Domain embeddings learned from the term collections

Domain	Lang	Size [MB]	Sentences	Tokens	Embedding voc
Agriculture	en	47.6	404,667	5,687,806	50,000
Anthropology	en	28.3	216,401	3,515,214	49,201
Architecture	en	44.6	337,175	5,547,847	50,000
Medicine	en	154.5	936,966	18,838,983	50,000
Music	en	3.9	28,285	492,824	10,311
Tourism	en	19.9	153,928	2,408,880	37,343
Agriculture	es	2.4	16,384	280,329	6304
Anthropology	es	4.0	16,384	280,329	9907
Architecture	es	2.9	22,132	355,369	7742
Medicine	es	5.4	34,254	651,945	12,335
Music	es	17.3	125,205	2,053,736	33,596
Tourism	es	1.3	10,801	155,216	3776

⁴³ <http://www.nltk.org/howto/wordnet.html>.

⁴⁴ <https://pypi.org/project/wikitoools/>.

Table 7 presents the global sizes of the datasets for all the domains. The figures correspond to the union of the collections extracted by the three systems. Looking at Fig. 9, we can see that WP corresponds to $B \cup C \cup D \cup E$, WN corresponds to $A \cup B \cup C$, $WP \cap WN$ corresponds to $B \cup C$, and WN_{all} corresponds to $B \cup C \cup D$.

In Table 8, we present the sizes in terms of the different datasets collected for our evaluation, as well as the sizes of the datasets collected by the other two systems. A bunch of rows presents the data for each dc . The first row in each bunch, identified as 'all', replicates the data of Table 7. Next rows show the data of the two baselines and our system. In the case of NG , we have included four versions having different threshold. NG_0 dataset contains all the data contained in NG ,⁴⁵ NG_4 , NG_5 and NG_6 set thresholds of 0.4, 0.5 and 0.6 on the PPR score. The last row in each bunch presents our results. Note that the figures of our system are slightly smaller than the ones shown in Table 11. The reason is that the comparison with NG , has to be performed only with terms attached at WP pages, categories are not considered.

Using the Extracted Terminology for Building Domain-Restricted Embeddings

In this section, we describe an indirect evaluation of our produced terminologies using them for learning accurate embeddings for each domain and language. These embeddings have been further evaluated and compared with several popular domain-independent ones.

Our intuition, supported by, for instance, [21, 45], or [47], is that for domain-restricted applications using domain-restricted embeddings, probably learned from smaller but more focused training sets, could be more effective than domain-independent embeddings, regardless the size of the training dataset.

Of course, learning embeddings from scratch using domain specific texts, as we do, is not the only approach for adapting embeddings to a domain. A popular alternative is *retrofitting* [11], that improves embeddings by enforcing the correlations with domain specific resources in a second training step. Another interesting approach is the use of monolingual dictionaries for learning embeddings constrained by maintaining the closeness between the embedding of a term and the embedding of its definition [5]. If the dictionary is terminological on the domain of interest this approach can be used for domain adaptation. [15] use concatenation of domain-independent and domain-restricted embeddings. The use of Transfer learning techniques [19] has obtained recently good results on transferring models between different languages and domains. None of these

approaches can be used here due to the lack of appropriate resources for all domains and both languages.

Building Domain-Restricted Embeddings

We have built embeddings for the two languages and the six domains of our evaluation set. For each language l and domain dc , the process is the following:

1. From our collection of terms, we select those corresponding to titles of WP categories, and from each WP category we obtain the set of WP pages belonging to such category. For the terms corresponding titles of WP pages, we collect the corresponding pages.
2. From the union of both sets, we collect the textual content of the pages.⁴⁶ This content is shallowly cleaned and constitutes the material used for learning the embeddings. In the Table 9, we show the main statistics of the texts used for learning the embeddings. Such table shows as well the size of the embedding vocabularies.
3. We learn the embeddings using the Facebook's *fastText* software⁴⁷ [20] and the Python wrapper to it, *fasttext*.⁴⁸ We have learned embeddings of size 200.
4. Domain-independent pre-trained embeddings have been used for comparison. The set of all the embeddings we have used for our evaluation is presented in Table 10. It is worth noting that only *MUSE*⁴⁹ domain-restricted embeddings have been learned from our data; in all the other cases, we have used pre-trained models downloaded from the different Websites.

Word2Vec and *Glove* are open domain resources, have been learned from big training data and their coverage are quite similar. *Pyysalo* has been learned from Medline collection for covering as possible the medical domain. These three embeddings are available only for English. From *MUSE*, we have included embeddings for English and Spanish, both pretrained open domain (rows 4 to 7 in the table) and domain restricted learned from our data (rows 8 to 19).

The column "size" refers to the size of the vocabulary, i.e. the terms for which embeddings vectors exist. It is worth noting that the size of our vocabularies are at least one order of magnitude smaller than the others.

⁴⁵ Tagged as 'accepted' in the original dataset.

⁴⁶ All the management of the WP pages and categories has been performed using the wikitools python library <https://github.com/alexzenwp/wikitools>.

⁴⁷ <https://fasttext.cc/>.

⁴⁸ <https://pypi.org/project/fasttext/>.

⁴⁹ Multilingual Unsupervised and Supervised Embeddings. See <https://research.fb.com/downloads/muse-multilingual-unsupervised-and-supervised-embeddings>.

Table 10 All embeddings used in our evaluation setting

Embedding	Domain	Lang	Url	References	Dim	Size
word2vec	none	en	https://radimrehurek.com/gensim/models/word2vec.html	[26]	300	3,000,000
Glove	none	en	https://nlp.stanford.edu/projects/glove/	[33]	200	400,000
Pyysalo	Medicine	en	http://bio.nlplab.org/	[34]	200	4,087,446
MUSE mono	none	en	https://github.com/facebookresearch/MUSEhttps://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md	[20]	300	50,000
MUSE mono	none	es			300	50,000
MUSE multi	none	en	https://github.com/facebookresearch/MUSE	[20]	300	50,000
MUSE multi	none	es			300	50,000
MUSE	Agriculture	en	https://github.com/facebookresearch/MUSE	[20]	200	50,000
MUSE	Anthropology	en			200	49,201
MUSE	Architecture	en			200	50,000
MUSE	Medicine	en			200	50,000
MUSE	Music	en			200	10,311
MUSE	Tourism	en			200	37,343
MUSE	Agriculture	es			200	6,304
MUSE	Anthropology	es			200	9,907
MUSE	Architecture	es			200	7,742
MUSE	Medicine	es			200	12,335
MUSE	Music	es			200	33,596
MUSE	Tourism	es			200	3,776

For evaluating the equality of the learned embeddings, and following [3] and [30], we have used two evaluation settings:

- Evaluating the embeddings as language models comparing for each domain the average similarities between a small set of terms of the domain and a set of out-of-domain terms.
- For the medical domain and the English language, we evaluate the embeddings using an analogy setting, i.e. the popular $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \sim \overrightarrow{queen}$, using as golden dataset the DDI corpus [18].

Evaluating the Embeddings as Language Models

We use the following metrics for this evaluation (for each language and domain):

- Degree of coverage of each dataset by each embedding. We compute it as the ratio between the number of terms

in the dataset also existing in the embedding vocabulary and the overall number of terms of the dataset.

- Degree of concentration of the terms of the domain dataset within the vectorial spaces of the embeddings. We compute this metric dividing the diameter of the subspace of embeddings of the terms of the domain dataset and the diameter of the embedding space. As these figures are costly to compute, we approximate them using small subsets of in-domain and out-of-domain terms. The process is detailed below.

For each domain dc and for each language l , we proceed as follows,

- We collect randomly a set of 20 terms existing in the WP for dc and l . Let $eval_{emb}^{l,dc}$ denote these sets.
- We collect for each language the intersection of the vocabularies of all the embeddings. After stopwords removing, we collect randomly from this set a set of the same size of out_of_domain terms. We denote as $eval_{emb}^{l,all}$ these two sets.
- For each embedding e , and using the metrics 3CosAdd and 3CosMul [30] we compute the average of distances

Table 11 Result of the basic evaluation (against WND) of terms compilation for some relevant domains. Values should interpreted according Figure 9

Domain Language	Tourism		Architecture		Music	
	EN	ES	EN	ES	EN	ES
Terms in WN						
Total (A+B+C)	556	180	219	30	1121	234
In WP (C)	6	10	7	4	18	144
Precision [%]	0.86	0.59	0.80	0.50	1.00	0.55
Recall [%]	1.00	1.00	0.89	1.00	1.00	1.00
New terms (D+E)	3,466	311	7,928	1,486	209	1373
Total new terms in the domain (C+D+E)	3,472	321	7,935	1,490	227	1517
Domain						
Language	Agriculture		Anthropology		Medicine	
	EN	ES	EN	ES	EN	ES
Terms in WN						
Total (A+B+C)	394	94	417	64	3468	512
In WP (C)	4	6	11	15	499	196
Precision [%]	0.67	0.55	0.79	0.54	0.78	0.55
Recall [%]	0.80	1.00	0.85	1.00	0.98	1.00
New terms (D+E)	394	510	2,294	1040	13,382	2,523
Total new terms in the domain (C+D+E)	398	516	2305	1055	13,881	2719

Table 12 Result of the evaluation of medical terms using Snomed-ct

Lang	EN	ES
Total terms	13,382	2523
Terms found	4195	994
Precision	31.3 %	39.4 %

between the members of $eval^{l,dc}$ and the average of distances between the members of $eval^{l,all}$.

- The averages are then compared for the different values of l , dc , and e .

Evaluating the Embeddings as Analogy

A popular way of evaluating embeddings is using them in an analogy task, i.e. “A is to B as C is to D”, for instance, “Spain is to Madrid as France is to X”, where X should have a vector very close to “Paris”. Most of the datasets used for evaluating embeddings within the analogy paradigm, as Country/Capital, Country/demonym, Verb/nominalisation, Verb_infinitive/verb_past, etc. can be easily built from Knowledge bases, as DBP, or linguistic patterns, i.e. following a distant learning approach. This is not the case of our domain-restricted setting.

Instead, we have devised the following setting: the evaluation is limited to the medical domain and the English language. We have evaluated in this setting the following embeddings: Word2Vec, Glove, Pyysalo, and the MUSE_en_medicine, learned from ours data.

We have used the Drug–Drug Interaction (DDI) corpus [48]. This corpus was developed for the *DDI Extraction 2013* challenge, [36]. It is made up of 792 texts selected from the DrugBank database, and 233 Medline abstracts. It is annotated with the mentions of drugs and their relations. Four types of relations are annotated in the corpus: ‘mechanism’, ‘effect’, ‘advise’, and ‘int’⁵⁰. We have chosen the ‘effect’ relation and the DrugBank subcorpus for our experiments, because there were the most frequent annotations, totalling 1855 pairs. We have filtered out from this dataset those pairs where none of the involved drugs occur in the union of the vocabularies of the embeddings to be evaluated. From this dataset of pairs $\langle D_1, D_2 \rangle$, we built a dataset of analogy quadruples $\langle D_1, D_2, D_3, D_4 \rangle$, i.e. D_1 effect D_2 as D_3 effect D_4 . The dataset of quadruples was randomly built and consists of 50,000 quadruples. We performed the evaluation using Gensim⁵¹ evaluation software, [35], able to process the four types of embeddings.

⁵⁰ For an undefined interaction.

⁵¹ <https://radimrehurek.com/gensim/>.

Results and Discussion

Results Obtained by Evaluation Against Magnini’s XWND

As mentioned in “Comparison with WND and NG Baseline Systems”, this evaluation was done using the terms already defined by Magnini and assuming their correctness. It is expected that terms discovered in *WP* will have similar precision values. The later consideration is also valid for all domains present in *MCR*.

Using the sets of terms as defined in Fig. 9, we calculate the corresponding precision/recall values shown in Table 11. For each language and domain, the initial number of *WN* variants and the precision/recall values are presented. As mentioned above, such values are calculated against information obtained from the Magnini’s domains. In the case of “Medicine”, for this particular domain there exist an external resource (*Snomed-ct*) that may be used as a reference list. Table 12 shows the results obtained in this case.

A first consideration to be taken into account in analysing the results shown in Table 11 is the specific characteristics of *WP* as a source of domain terms. In particular:

1. There exist differences in the category graph among different languages. See for example the domains “Medicine” and “Veterinary”. Although definitions are similar in both Spanish and English *WPs*, the former considers “Veterinary” at the same level of “Medicine” whilst the latter considers it as a subcategory of “Medicine”. This difference causes a main difference in the *TC* direct/indirect linked to them;
2. English *WP* is a resource densely-linked; this fact may cause unexpected relations among *TC* and domains. Consider for example the domain “Agriculture” and the terms “abdomen” or “aorta”. Due to a link among “Agriculture” → “veterinary medicine”, such candidates are considered as belonging to the domain which may be considered correct or not depending of the view point;
3. *WP* is an encyclopaedic resource updated by heterogeneous people and its final use is also heterogeneous, therefore the termhood of some *TC* may be controversial. See for example: “list of architecture topics” in “Architecture” (English) or “Academia Panamericana de Anatomía” (Pan American Academy of Anatomy) in Medicine (Spanish). Such strings are closely related to their corresponding domains but they do not have any terminological value. Consequently, they should be filtered out but this is not always an easy task.
4. The use of *Snomed-ct* to evaluate terms recovered in “Medicine” allows a better evaluation (at least it allows to use a more complete reference set). But it should also

Table 13 Examples of validated and non-validated medical terms using Snomed-ct

Lang	EN	ES
Valid/true	Fibrosarcoma	bronquiolitis
	Kirschner wire	duodeno
	Pneumaturia	placa dental
Valid/false	Alzheimer research forum	especialidades médicas
	Birmingham accident hospital	
Invalid/true	Rivalta test	otitis externa
		circulación portal hepática
		ortesis
Invalid/false	High-throughput screening	condenado a vivir
	Kawasaki medical school	asociación pablo ugarte
	Goldwater rule	huérfano del sida

be taken into consideration that *Snomed-ct* is defined as “a standardised, multilingual vocabulary of clinical terminology that is used by physicians and other health care providers for the electronic exchange of clinical health information”.⁵² Therefore, it does not include all the terminology of the domain but only those terms commonly used in the professional daily practice. This fact is the cause of some false error indication such as:

- (a) Complex term: Some terms in this database are coordinated terms. See for ex-ample the Spanish TC: *enfermedades hereditarias y degenerativas del sistema nervioso central* (genetic and degenerative disorders of the central nervous system). It causes that none of the coordinated term are detected.
- (b) Some entries exist only as specialised. See for example the Spanish TC glándula (gland), it only exists as a more specialised terms like *glándula esofágica* (esophageal gland) or *glándula lagrimal* (lacrimal gland).
- (c) Missing terms like: *andrología* (andrology) or *arteria cerebelosa media* (medial cerebellar artery), present only in *WP* snapshot used for this experiment. Snomed-ct is a continuous evolving resource that for some reason do not include absolutely all terms used in everyday medicine. This fact reflects in the low precision values showed in Table 12.

Table 13 shows some examples about what our filtering system consider as valid or invalid terms and the actual status of such terms. See for example the term *fibrosarcoma*, it has has be considered correctly as a valid term. But the string *alzheimer research forum* has been also considered a term but in spite it relation to the domain it has no terminological value.

The reverse example is represented by the string *kawasaki medical school* that has been correctly rejected by our system. This examples shows the difficulty in effectively filtering terms that are closely related to the area but as a matter of fact do not have any terminological value.

5. The precision and recall values shown in Table 11 for “Medicine” are influenced by the fact the WN used in this task has been locally improved by about 2000 new synsets not present in the standard WN.

Comparison with NG Baseline Results

The results of the comparison with the *NG* baseline system are presented in Table 14. The usual measures of precision, recall and F1 (uniform weighted harmonic mean of precision and recall) are applied only on the terms occurring not only in *WP* but also in *WN* ($WP \cap WN$) for which a golden set exists. For the general case of *WP* terms, our guess is that these measures should hold but the only fair measure we can use is the coverage.

As can be seen in Table 14, our system consistently outperforms *NG* at all the threshold levels both within $WP \cap WN$, for which *XWND* acts are golden set, in terms of F1 for all the domains and within *WP* in terms of coverage for all the domains but “Music”. The results are based on our advantage on recall while our precision are slightly below than *NG* ones.

The problematic case is “Music”, due to an extremely low coverage, due to a low number of categories selected that resulted on a low number of pages (139, as shown in Table 5). The detailed results of our experiments are presented in Table 15. It is worth noting that the true evaluation is performed only for terms in $WP \cap WN$, i.e. domains coded with *WND* existing in the *MCR* and in *WP*. For this measure Magnini’s has advantage over *NG* and our system, both in term of precision and recall, and, thus, in F1 because the gold standard is Magnini’s *WN* dataset and most of the terms in *WN* occur also in *WP*, so, for Magnini $WP \cap WN$ and *WN* are basically the same.

⁵² <https://searchhealthit.techtarget.com/definition/SNOMED-CT>.

Table 14 Results of the comparison with baseline systems

dc	prec	Recall	f1	cov	prec	Recall	f1	cov
Agriculture					Anthropology			
ALL	0.93	1	0.96	1	0.95	1	0.98	1
MAG-NINI	0.46	0.20	0.28	0.06	0.38	0.34	0.36	0.29
NG0	1	0.17	0.29	0.03	1	0.2	0.33	0.07
NG4	1	0.17	0.29	0.03	1	0.20	0.33	0.07
NG5	1	0.13	0.24	0.02	1	0.18	0.30	0.06
NG6	1	0.03	0.07	0.01	1	0.06	0.11	0.02
OURS	0.8	0.88	0.84	0.95	0.87	0.64	0.74	0.72
Architecture					Medicine			
ALL	0.98	1	0.99	1	0.94	1	0.97	1
MAG-NINI	0.02	0.02	0.03	0.40	0.23	0.24	0.24	0.47
NG0	1	0.20	0.35	0.09	1	0.22	0.3	0.11
NG4	1	0.20	0.35	0.09	1	0.22	0.36	0.11
NG5	1	0.17	0.30	0.08	1	0.19	0.32	0.1
NG6	1	0.05	0.11	0.02	1	0.07	0.12	0.03
OURS	0.88	0.59	0.71	0.60	0.87	0.69	0.77	0.58
Music					Tourism			
ALL	0.92	1	0.96	1	0.94	1	0.97	1
MAG-NINI	0.54	0.57	0.59	0.91	0.33	0.33	0.33	0.28
NG0	1	0.19	0.32	0.18	1	0.21	0.34	0.06
NG4	1	0.19	0.32	0.18	1	0.21	0.34	0.06
NG5	1	0.17	0.29	0.16	1	0.17	0.3	0.05
NG6	1	0.07	0.13	0.07	1	0.06	0.11	0.02
OURS	1	0.62	0.77	0.10	0.92	0.66	0.77	0.73

Results of the Domain-Restricted Embeddings

As already mentioned in “Using the extracted terminology for building domain-restricted embeddings”, we have devised two evaluation settings for our domain restricted embeddings. We present the results in the following two subsections.

Results of the Evaluation of the Embeddings as Language Models

The results are presented in Tables 16, 17 and 18. In Table 16, we show the degree of coverage. Rows 1 and 2 show the sizes of the English and Spanish datasets for the 6 domains. Rows from 3 to 7 present the results of the open domain embeddings for English⁵³, rows from 8 to 13, the domain embeddings for English. The rest of the table presents the results for Spanish.

Looking at the table some facts can be seen :

- Between the out of domain embeddings, for English, *Word2Vec* (row 6) presents the best absolute results in all the domains but “Medicine”. The distribution between domains is coherent.
- As expected, *Pyysalo* provides the best results on “Medicine”
- When looking to the weighted figures, i.e. dividing the counts in Table 16 by the corresponding sizes of embedding vocabularies (last column of Table 16) the best scored embeddings are for each domain and language are the *MUSE* embeddings learned with our data⁵⁴ These scores are highlighted in Table 16. It is worth noting that besides obtaining the best weighted scores in all the domains our system obtained the best absolute scores in two domains: Anthropology and Tourism.

The results for the second measure are presented for English in Table 17 and for Spanish in Table 18. Remember that the figures in both tables show the ratio between the average

⁵³ We have included, though, in this group *Pyysalo* embeddings belonging to the Medical domain.

⁵⁴ With the only exception of “Tourism” for Spanish, due to the reasons already discussed.

Table 15 Overall results of the evaluation over baseline systems

sys	WP	WN	$WP \cap WN$	WNall	prec	Recall	f1	cov
ALL	206,413	27,575	27,566	117,311	0.36	1	0.50	1
MAG	25,849	3765	25,848	25,848	1	0.94	0.96	0.20
NG0	151,176	–	4737	9,4482	0.10	0.17	0.12	0.64
NG4	151,176	–	4737	9,4482	0.10	0.17	0.12	0.64
NG5	119,364	–	4122	78,192	0.11	0.15	0.11	0.51
NG6	30,270	–	1371	21,834	0.13	0.05	0.06	0.13
OURS	35,633	–	6106	6259	0.99	0.21	0.32	0.21

distances of the embeddings of the domain datasets and the average distances of the embeddings of the out-of-domain dataset. All the datasets have a size of 20 terms. The lower this ratio is the better is the embedding. The ideal situation is when the ratio is below 1. The conclusions are basically the same for the two languages and can be summarised as follows:

- The out-of-domain embeddings show in general a low performance. *Word2Vec*, for instance have scores higher than 1.9 for all the domains.
- As in the case of the first measure *Pyysalo* offers a good result on “Medicine”, but, surprisingly, not the best.
- An unexpected case is *Glove* that presents the best results in two domains, including “Medicine”, and the second in another one.
- The domain dependent *MUSE* embeddings learned with our data obtain the best overall performance with two first positions (“Architecture” and “Music”), two second positions (“Anthropology” and “Tourism”) and a third position in “Medicine”.

If we combine the two measures our embeddings get consistently the best scores for all the domains and the two languages.

Results of the Evaluation of the Embeddings in the Analogy Task

Table 19 presents the results of the evaluation of the Analogy task. Column 2 shows the absolute counts, i.e. the number of correct relations discovered by analogy for the different systems. Remember that the domain is “Medicine” and the language English. In column 4, we present the results weighted by the size of the embeddings. Once again *Glove* obtain the best results in absolute value but our system surpasses it when using the weighted scores for getting the best results. When looking at the table it could be striking the extremely low values of the ratios in column 4. The reason is that, for the sake of comparability of the embeddings, we have weighted the results by dividing them by the size of the

resource. The new figures look a bit odd but provide a more fair material for comparison.

Conclusions

In this paper, we have presented a new approach for compiling a multi-domain multi-language terminology. For obtaining such resource, we use the information obtained from three main resources: the *WP* encyclopedia, the *MCR* knowledge base and the cross-domain ontology *DBpedia*. The resources have been compiled for English and Spanish and for the full set of 164 domains as defined in [23] and [4]. The terms collected has been formatted according OLIF, an XML interchange format specifically designed lexical/terminological data exchange. The collected terminologies, consisting of titles of either pages or categories of the corresponding *WP*, contain globally 472,685 English terms and 162,842 Spanish terms, 79,946 of which are cross-linguistically mapped. These figures clearly point out the importance of the resource.

The results were evaluated in several ways. For the first scenario, we perform a partial evaluation restricted to the terms occurring in both *MCR* and *DBpedia*. Its main advantage is that can be applied to any combination of domain and language. For those few domains where an external reliable reference is available a more complete evaluation scenario was foreseen. This was the case of “Medicine” for which we use *Snomed-ct*, a well-known medical term repository.

We also perform an indirect evaluation of the resulting terminologies by using them for learning accurate word embeddings for each domain and language. We produce such embeddings for six domains and both English and Spanish languages. These embeddings have been further evaluated using two settings. The first one comparing with several popular domain-independent ones and a second one specifically for medicine using as golden dataset the DDI corpus.

The results of these evaluations have been positive and suppose a warranty of the quality of the resource.

In any case and as foreseen the results, evaluation is a difficult task, mainly due to issues in the reference list even in the case a more complete term reference list can be used.

Table 16 Evaluation of embeddings by coverage of the terms datasets

emb	Agriculture	Anthropology	Architecture	Medicine	Music	Tourism	Total
Terms en	8640	2258	7852	13,243	206	3437	35,636
Terms es	500	1012	1453	2455	1290	302	7012
museBig_en_open	9	3	9	22	0	4	47
museMult_en_open	11	6	8	20	2	3	50
pyysalo_en_medicine	273	173	283	2370	23	153	3275
word2vec_en_open	430	194	525	1,313	31	281	2774
glove_en_open	58	22	114	131	4	37	366
muse_en_anthropology	20	227	27	16	8	45	343
muse_en_architecture	46	33	366	17	4	46	512
muse_en_tourism	15	46	57	15	6	326	465
muse_en_medicine	7	14	14	428	5	9	477
muse_en_music	2	7	4	6	24	5	48
muse_en_agriculture	186	28	63	43	3	25	348
museBig_es_open	9	4	9	23	0	4	49
museMult_es_open	7	5	6	28	1	5	52
muse_es_anthropology	0	46	5	0	4	4	59
muse_es_architecture	1	1	53	1	2	9	67
muse_es_tourism	0	0	6	0	20	0	26
muse_es_medicine	0	0	1	45	4	1	51
muse_es_music	1	2	3	4	352	2	364
muse_es_agriculture	17	1	1	0	1	1	21

Table 17 English embeddings evaluation as language model

System	Agriculture	Anthropology	Architecture	Medicine	Music	Tourism
<i>muse</i> _{en_agriculture}	1.116	1.176	1.025	1.587	1.223	1.389
<i>muse</i> _{en_anthropology}	1.683	1.152	1.308	1.971	1.543	1.583
<i>muse</i> _{en_architecture}	1.600	1.172	0.986	2.112	1.461	1.318
<i>muse</i> _{en_medicine}	1.682	1.456	1.393	1.151	1.679	1.501
<i>muse</i> _{en_music}	0.665	0.817	1.021	1.104	0.748	0.999
<i>muse</i> _{en_tourism}	1.703	1.119	1.005	1.913	1.309	1.139
<i>glove</i> _{en_open}	0.966	0.618	1.232	0.891	1.121	1.012
<i>museBig</i> _{en_open}	1.935	1.139	1.313	1.946	1.862	1.489
<i>museMult</i> _{en_open}	1.935	1.139	1.313	1.946	1.862	1.489
<i>pyysalo</i> _{en_medicine}	2.560	2.337	2.487	0.897	2.446	2.520
<i>word2vec</i> _{en_open}	2.577	1.974	2.047	3.215	3.302	2.733

Bold values indicate degree of concentration of the terms of the domain dataset within the vectorial spaces for each domain code and embedding

Also the encyclopaedic character of *WP* conditioned the list of new terms obtained. As foreseen, due to the different completeness of the used resources, the performance may change according the domain/language considered.

In the future, we plan to improve the exploration of the *DBpedia* to reduce the number of false domain terms by improving the detection of neutral categories. Other possibilities for improving the final list of terms are: (i) using the *WP* article text as another factor about the pertinence of a given page; (ii) a better integration of both bottom-up and

top-down exploration procedures. We also plan to enlarge the number of proposed *TC* by using interwiki information.

Another evident line of research is the application of the system to other languages. Taking into account the limited use of linguistic processors and the wide availability of the involved resources for a large set of languages, our guess is that the system could be applied successfully to many other languages. As a first step in this line, we are working on the application of the system to Arabic, Basque, Catalan, French and German.

Table 18 Spanish Embeddings evaluation as language model

System	Agriculture	Anthropology	Architecture	Medicine	Music	Tourism
<i>muse</i> _{es_agriculture}	0.994	0.962	1.120	1.127	1.298	0.970
<i>muse</i> _{es_anthropology}	1.145	1.122	1.159	1.436	1.108	1.120
<i>muse</i> _{es_architecture}	1.230	1.066	1.071	1.484	1.044	0.937
<i>muse</i> _{es_medicine}	1.292	1.013	1.064	0.747	1.255	1.002
<i>muse</i> _{es_music}	1.208	1.467	1.399	1.612	1.202	1.481
<i>muse</i> _{es_tourism}	1.252	1.103	1.019	1.023	1.143	1.0234
<i>museBig</i> _{es_open}	1.602	1.085	1.058	1.661	1.268	1.208
<i>museMult</i> _{es_open}	1.602	1.085	1.058	1.661	1.268	1.2087

Bold values indicate degree of concentration of the terms of the domain dataset within the vectorial spaces for each domain code and embedding

Table 19 Evaluation of medical embeddings using DDI corpus

System	Relations	Size	Ratio
<i>muse_en_medicine</i>	86	12,335	0.006972
<i>glove_en_open</i>	266	400,000	0.000665
<i>word2vec_en_open</i>	131	3,000,000	0.0000437
<i>pyysalo_en_medicine</i>	154	4,087,446	0.0000377

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest 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).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Agirre E, Soroa A. Personalizing pagerank for word sense disambiguation. In: Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009). Athens, Greece; 2009.
- Aronson AR, Lang F-M. An overview of metamap: historical perspective. J Am Med Inform Assoc (JAMIA). 2010;17(3):229–36.
- Bakarov A. A survey of word embeddings evaluation methods; 2018. *CoRR*, [arXiv:1801.09536](https://arxiv.org/abs/1801.09536).
- Bentivogli L, Forner P, Magnini B, Pianta E. Revising the wordnet domains hierarchy: Semantics, coverage and balancing. In: Proceedings of the Workshop on Multilingual Linguistic Resources, MLR'04, pp. 101–108, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- Bosc T, Vincent P. Auto-encoding dictionary definitions into consistent word embeddings. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp 1522–1532. Association for Computational Linguistics, 2018.
- Cabrera-Diego LA, Sierra G, Vivaldi J, Pozzi M. Using wikipedia to validate the terminology found in a corpus of basic textbooks. In: Chair NCC, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S, editors, Proceedings of the eight international conference on language resources and evaluation (LREC'12), pp. 3820–3827, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).
- Cabré MT, Estopà R, Vivaldi J. Automatic term detection: a review of current systems. In: Recent advances in computational terminology, pp 53–87. John Benjamins, 2001.
- Conrado MS, Pardo TAS, Rezende SO. Exploration of a rich feature set for automatic term extraction. In: Castro F, Gelbukh A, González M, editors. Advances in artificial intelligence and its applications. Berlin: Springer; 2013. p. 342–54.
- Drouin P. Term extraction using non-technical corpora as a point of leverage. Terminology. 2003;9(1):99–115.
- Enguehard C, Pantera L. Automatic natural acquisition of a terminology. J Quant Linguist. 1995;2(1):27–32.
- Faruqui M, Dodge J, Jauhar SK, Dyer C, Hovy E, Smith NA. Retrofitting word vectors to semantic lexicons. In: Proceedings of NAACL; 2015.
- Fellbaum C. Wordnet: an electronic lexical database. Libr Q. 1999;69(3):406–8.
- Fernández-Montraveta A, Vázquez G, Fellbaum C. The Spanish version of wordnet 3.0. In: Storrer A, Geyken A, Siebert A, Würzner K-M, editors. Text resources and lexical knowledge. Berlin: Mouton de Gruyter; 2008. p. 175–82.
- Frantzi KT, Ananiadou S, Tsujii J. The c-value/nc-value method of automatic recognition for multi-word terms. In: Nikolaou C, Stephanidis C, editors. Research and advanced technology for digital libraries. Berlin: Springer; 1998. p. 585–604.
- Goikoetxea J, Agirre E, Soroa A. Single or multiple? Combining word representations independently learned from text and wordnet. In: Schuurmans D, Wellman MP, editors, Proceedings of the Thirtieth AAAI conference on artificial intelligence, February 12–17, 2016, Phoenix, Arizona, USA. AAAI Press; 2016. p. 2608–2614.

16. González-Agirre A, Castillo M, Rigau G. A proposal for improving wordnet domains. In: Chair NCC, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S, editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).
17. Heid U, Jauß S, Krüger K, Hohmann A. Term extraction with standard tools for corpus exploration. Experience from German. In: Proceedings of TKE'96. Berlin: Springer; 1996, p. 255–279.
18. Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. *J Biomed Inform.* 2013;46(5):914–20.
19. Kale M, Siddhant A, Nag S, Parik R, Grabmair M, Tomasic A. Supervised contextual embeddings for transfer learning in natural language processing tasks; 2019.
20. Lample G, Conneau A, Ranzato M, Denoyer L, Jégou H. Word translation without parallel data. In: 6th International conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, conference track proceedings. OpenReview.net; 2018.
21. Limsopatham N, Collier N. Modelling the combination of generic and target domain embeddings in a convolutional neural network for sentence classification. In: Proceedings of the 15th workshop on biomedical natural language processing. Association for Computational Linguistics, 2016.
22. Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M. Biomedical term extraction: overview and a new methodology. *Inf Retr J.* 2016;19(1):59–99.
23. Magnini B, Cavaglià G. Integrating subject field codes into wordnet. In: Proceedings of the language resources and evaluation conference (LREC 2000), Athens, Greece, May 2000. ELRA, p. 1413–1418.
24. Medelyan O, Milne D, Legg C, Witten IH. Mining meaning from wikipedia. *Int J Hum-Comput Stud.* 2009;67(9):716–54.
25. Mihalcea R, Tarau P, Figa E. Pagerank on semantic networks, with application to word sense disambiguation. In: COLING 2004, 20th international conference on computational linguistics, proceedings of the conference, 23–27 August 2004, Geneva, Switzerland; 2004.
26. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Bengio Y, LeCun Y, editors, 1st international conference on learning representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, workshop track proceedings; 2013.
27. Newman D, Koilada N, Lau JH, Baldwin T. Bayesian text segmentation for index term identification and keyphrase extraction. In: Proceedings of COLING 2012, Mumbai, India. The COLING 2012 Organizing Committee; 2012, p. 2077–92.
28. Ni CS, Jiang JC. Degree, closeness, and betweenness: application of group centrality measurements to explore macro-disciplinary evolution diachronically. In: Proceedings of the ISSI; 2011, p. 1–13.
29. Niemann E, Gurevych I. The people's web meets linguistic knowledge: automatic sense alignment of wikipedia and wordnet. In: Proceedings of the 9th international conference on computational semantics; 2011, p. 205–14.
30. Omer Levy YG. Linguistic regularities in sparse and explicit word representations. In: Proceedings of the thirteenth conference on computational natural language learning; 2014, p. 171–80.
31. Page L, Brin S, Motwani R, Winograd T. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November (1999). Previous number = SIDL-WP-1999-0120.
32. Pazienza MT, Pennacchiotti M, Zanzotto FM. Terminology extraction: an analysis of linguistic and statistical approaches. In: Sirmakessis S, editor. Knowledge mining. Berlin: Springer; 2005. p. 255–79.
33. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: Empirical methods in natural language processing (EMNLP); 2014, p. 1532–43.
34. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional semantics resources for biomedical text processing. In: LBM; 2013, p. 1532–43.
35. Řehůřek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks, Valletta, Malta. ELRA; 2010, p. 45–50. <http://is.muni.cz/publication/884893/en>.
36. Segura-Bedmar I, Martínez P, Zazo MH. Lessons learnt from the DDI extraction-2013 shared task. *J Biomed Inform.* 2014;51:152–64.
37. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE J Biomed Health Inform.* 2018;22(5):1589–604.
38. Spackman K, Campbell K, Cote R. Snomed rt: a reference terminology for health care. In: Proceeding of AMIA annual fall symposium, vol. 4; 1997, p. 640–4, 02.
39. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. In: Proceedings of the 27th international conference on artificial neural networks (ICANN 2018); 2018.
40. Vivaldi J. Extracción de candidatos a término mediante combinación de estrategias heterogéneas. PhD thesis, Polytechnical University of Catalonia, 6 2001.
41. Vivaldi J, Rodríguez H. Using wikipedia for term extraction in the biomedical domain: first experience. *Procesamiento del Lenguaje Natural.* 2011;45:251–4.
42. Vivaldi J, Rodríguez H. Using wikipedia for scaling up domain terms extraction in the humanities. In: Proceedings of the second workshop on creation, harmonisation and application of terminology resources, Madrid, Spain; 2012.
43. Wang R, Liu W, McDonald C. Featureless domain-specific term extraction with minimal labelled data. In: Proceedings of the Australasian language technology association workshop 2016, Melbourne, Australia; 2016, p. 103–12.
44. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le Pon AD, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 2017;46(D1):D1074–82.
45. Xu C, Bai Y, Bian J, Gao B, Wang G, Liu X, Liu T-Y. Rc-net: A general framework for incorporating knowledge into word representations. In: CIKM; 2014, p. 1219–28.
46. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. In: CoRR. [arXiv:1708.02709](https://arxiv.org/abs/1708.02709), 2017.
47. Yu M, Dredze M. Improving lexical embeddings with semantic knowledge. In: ACL; 2014, p. 545–50.
48. Zazo MH, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. *J Biomed Inform.* 2013;46(5):914–20.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.