Knowledge and Information Systems (2023) 65:1365–1397
https://doi.org/10.1007/s10115-022-01767-5

REGULAR PAPER

Tailoring and evaluating the Wikipedia for in-domain comparable corpora extraction

Cristina España-Bonet[1] · Alberto Barrón-Cedeño[2] · Lluís Màrquez[3]

Received: 17 June 2020 / Revised: 9 September 2022 / Accepted: 10 September 2022 /
Published online: 1 November 2022

Abstract
We propose a language-independent graph-based method to build *à-la-carte* article collections on user-defined domains from the Wikipedia. The core model is based on the exploration of the encyclopedia's category graph and can produce both mono- and multilingual comparable collections. We run thorough experiments to assess the quality of the obtained corpora in 10 languages and 743 domains. According to an extensive manual evaluation, our graph model reaches an average precision of 84% on in-domain articles, outperforming an alternative model based on information retrieval techniques. As manual evaluations are costly, we introduce the concept of *domainness* and design several automatic metrics to account for the quality of the collections. Our best metric for domainness shows a strong correlation with human judgments, representing a reasonable automatic alternative to assess the quality of domain-specific corpora. We release the `WikiTailor` toolkit with the implementation of the extraction methods, the evaluation measures and several utilities.

Keywords Comparable corpora · Wikipedia category graph · Domain-specific corpora · Domainness metrics

1 Introduction

Different natural language processing (NLP) and information retrieval (IR) tasks require large amounts of domain-specific text with different levels of parallelism. With such data, one can obtain in-domain lexicons, semantic representations of concepts, train specialized machine

✉ Cristina España-Bonet
cristinae@dfki.de; cristinae.uni@gmail.com

Alberto Barrón-Cedeño
a.barron@unibo.it

Lluís Màrquez
lluismv@amazon.com

[1] DFKI GmbH, Saarbrüken, Germany

[2] Università di Bologna, Forlì, Italy

[3] Amazon, AWS AI Labs, Barcelona, Spain

translation engines or question answering systems. A common strategy to gather multilingual domain-specific material is crawling the Web; e.g., looking for different language editions of a website [18, 43]. Nowadays, one of the largest controlled sources for this kind of text at the fingertips is the Wikipedia—an online encyclopedia with millions of topic-aligned articles in multiple languages.[1]

In this article, we explore the value of the Wikipedia as a source for domain-specific comparable text with a practical perspective. Our contributions follow two directions. From a theoretical point of view, we introduce:

1. *A novel methodology for article selection.* We extract in-domain articles taking advantage of Wikipedia's densely connected category graph. The multilingual aspect of the resource facilitates the extraction of cross-language counterparts.
2. *A novel concept to assess the quality of an in-domain collection.* We define *domainness* as a combination of the representativity and cohesion of texts and introduce several automatic metrics that model both. The correlation between our metrics and a manual evaluation allows us to validate the metrics which reduce the necessity of relying on expensive manual evaluations in the future.

From a pragmatic point of view, we release:

3 *An open-source software implementation of our architectures and quality metrics.* `WikiTailor` is a Java toolkit designed to extract and analyze corpora from Wikipedia in any language and domain.[2] `WikiTailor` makes obtaining multilingual in-domain data from the Wikipedia easy.

4 *The corpora derived from our experiments.* We make available the collections obtained with our best models and the domain-specific term vocabularies for 743 domains in 10 languages: English, French, Spanish, German, Arabic, Romanian, Catalan, Basque, Greek, and Occitan [17].

The rest of the paper is distributed as follows. Section 2 overviews comparable corpora acquisition methods, with special focus on the categorization and multilinguality of the Wikipedia, the relevance of Wikipedia for NLP and IR, and related work. Section 3 presents our models for the automatic extraction of (multilingual) in-domain corpora. Section 4 describes the experimental settings, analyses the characteristics of the collections extracted, and reports the results of our manual evaluation to assess their quality. In Sect. 5, we define the concept of domainness and introduce several automatic evaluation metrics. In Sect. 6, we use them to quantify the quality of the collections produced. We draw conclusions in Sect. 7. Appendix A contains a glossary with Wikipedia-specific terms, whereas Appendix B summarizes the input parameters accepted by `WikiTailor`. Appendix C offers further details of the crowdsourcing experiment that leads to our manual evaluation.

2 Comparable corpora and the Wikipedia

Multiple kinds of Web contents have been used as a source for the acquisition of comparable corpora. Usually, the process involves two steps. First, documents in the required languages are acquired [2, 38, 43, 51]. Second, an alignment identifies pairs of comparable documents [21, 34, 41, 52, 54]. Among these works, [38] and [21] are specially relevant, since their

[1] http://www.wikipedia.org, with 314 active languages in December 2021.

[2] A stand-alone executable and the source code are available at http://cristinae.github.io/WikiTailor.

corpus is Wikipedia. In this case, and up to the limitations we discuss later, alignment is close to trivial, thanks to the existing links across articles in different languages.

Three properties make the Wikipedia a particularly suitable source of comparable and parallel data: (*i*) it has editions in a large number of languages; (*ii*) articles covering the same topic in different languages are connected via *inter-language links*, also called *langlinks*; and (*iii*) articles have categories which purpose is both describing the topic covered and grouping together related articles. Nevertheless, it also has drawbacks. (*i*) The inter-language links are subject to inconsistencies because, in general, they are manually created by volunteers. Not only could volunteers make mistakes linking non-equivalent concepts, there are articles in one edition that are connected to more than one article in another language [27]. (*ii*) An article can belong to multiple categories. Indeed, it is possible to construct loops with categories; i.e., a non-strict tree hierarchy is in place [57]. (*iii*) Given that categories are built collaboratively, they are often arbitrary. Many articles lack a proper association with the categories they should belong to, and there is an over-categorization phenomenon.[3] Consequently, the Wikipedia category graph (WCG) and the inter-language links must be used carefully when extracting domain-aligned articles across multiple language editions. Moreover, the intersection of common articles across languages tends to be small. In general, smaller Wikipedia editions are not subsets of the largest ones. In the dumps considered for this study, only 0.4% of the articles are common across the ten languages with all ten within the top-100 Wikipedia editions in terms of size. For the largest four editions (English, French, Spanish, German), representing relatively close cultures, the number grows to 4.8% only. This is the so-called *context diversity* effect [27]: the articles in the intersection correspond to globally relevant concepts, whereas the singletons represent cultural diversity. We use the globally relevant concepts when selecting our domains of study, as we expect them to have the most comparable articles.

The Wikipedia has been widely and successfully used in (CL)-NLP and (CL)-IR. For example, it has been used for terminology and bilingual dictionary extraction [11, 16, 28, 42, 56]. Wikipedia's inter-language links are crucial to obtain an aligned comparable corpus. The value of the Wikipedia as a source of highly comparable and parallel sentences has been appreciated over the years [1, 5, 9, 37, 47–49, 55]. With the rise of deep learning for NLP and the need of large amounts of *clean* data, the use of Wikipedia has grown exponentially not only for parallel sentence extraction and machine translation [25, 44, 46, 53], but also for semantics. Word and contextual embeddings have been trained on it and made available for more than 100 languages. Examples include fastText [6, 24] and MUSE word embeddings [30], multilingual BERT [15], and LASER sentence embeddings [3]. Newer and larger models trained on orders of magnitude more data, such as GPT-3 [8], mT6 [10] and ΔLM [33], include Wikipedia in the training dataset.

Semantic representations can also be obtained via explicit semantic analysis (ESA) [20] and have been widely used in IR to compute the semantic relatedness of concept vectors. CL-ESA [26, 40] is a cross-language extension which allows for computing semantic relatedness across languages. Compared to neural network embeddings, CL-ESA representations are less sensitive to the amount of training data and differences in sizes among languages (see Sect. 5). Therefore, they are adequate within the multilingual setting we present in this work.

A number of efforts have been focused on producing comparable collections from the Wikipedia. The authors of [21] proposed the basis to exploit the metadata (category tags)

[3] This is stressed in the Wikipedia itself: http://en.wikipedia.org/wiki/Wikipedia:Overcategorization; last visited: December 2021.

and the WCG to extract different comparable subsets. They distinguished three kinds of collections. (*i*) Non-aligned: articles belonging to the same topic just because they have the same associated category; (*ii*) strongly aligned: articles connected through an inter-language link, both belonging to the same category; (*iii*) softly aligned: articles connected by an inter-language link but not necessarily belonging to the same category. Their CorpusPedia tool[4] extracts comparable corpora from the Wikipedia having as input a pair of languages and a category. Our alignment is of the first type. We go beyond and deal with complete domains rather than with individual categories. We extract domains exploring the WCG; we extract more articles by avoiding their "strict" strategy based only on the exact category and its children. This idea was first sketched in [5], where we also extracted parallel sentences from the identified comparable corpora in *Computer science*, *Science* and *Sport* to domain-adapt a machine translation system.

The WCG is close to a taxonomy structure [57]. Still, exploring it might be slow given the size of some Wikipedia editions and the high density of their graph with numerous loops. Several works facilitate the task. In PetScan[5], a user inputs one or more categories and gets all their associated articles up to a desired depth. [4] introduced a graph database structure and provided a database for the English Wikipedia with monthly updates. Graph databases have the advantage of allowing traversing and performing breadth-first search efficiently. Different to us, all these utilities expect the user to input the depth up to which define the traversal for a root category.

In an approach completely unrelated to graphs, the authors in [37] and [38] proposed a model based on a typical search engine. Given two Wikipedia editions in language L and L', they identify the subset of article pairs in L and L' (i.e., connected by an inter-language link) and index the resulting documents. The index is queried with the most frequent 100 keywords from an external in-domain corpus to retrieve the relevant articles. The information about the Wikipedia structure is neglected, and the selection of in-domain articles fully depends on their contents. Due to the completely different nature of this system with respect to our approach, we adopt it for comparison purposes. In [37] the authors also showed the difficulties of using Wikipedia categories for the extraction of articles in the *Alpine* domain. They found that some articles within the main namespace lack a category tag and that the categories assigned to the same article in different languages do not overlap.

Full projects have been devoted to the topic. ACCURAT[6] implemented a toolkit for alignment and information extraction from comparable corpora but, unfortunately, it is not available. The toolkit [36] performs alignment of comparable documents, extraction of parallel sentences, extraction of terminologies, and extraction of named entities. It can be applied on the Wikipedia to extract a general domain comparable corpus and retrieves the documents by analyzing comparable segments in the candidates. A series of similarity metrics is applied to determine the level of comparability between two documents. The approach and aim of the tool is completely different to ours. They focus on comparability, regardless of the domain. Our focus is the domain, and the comparability is a direct consequence: at the corpus level, if the languages cover the same domain, the corpora are comparable and at the document level, comparability can be established using the inter-language links.

Linguatools[7] released three corpora derived from the Wikipedia in 23 languages: a monolingual corpus with more than 5 billion tokens, a comparable corpus with more than 41

[4] http://gramatica.usc.es/pln/tools/CorpusPedia.html.

[5] https://petscan.wmflabs.org/.

[6] http://www.accurat-project.eu/.

[7] http://linguatools.org/.

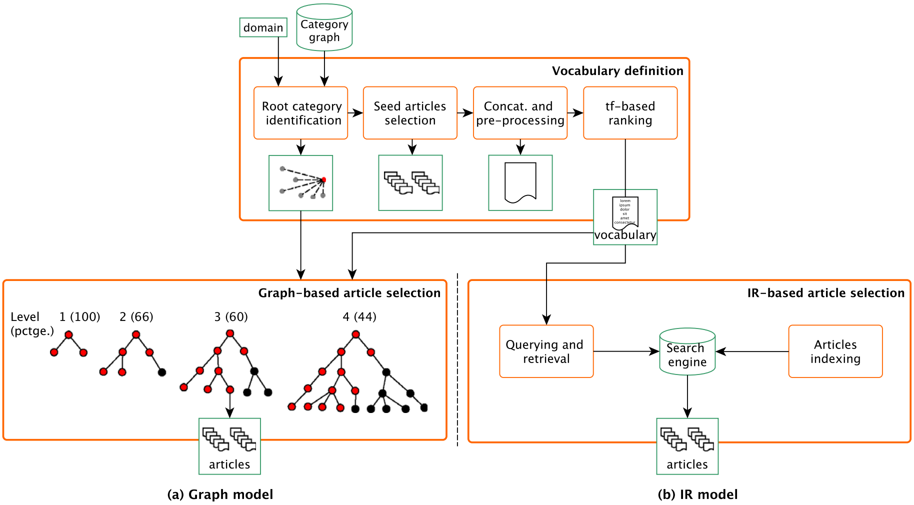


Fig. 1 Domain article selection pipelines. Both pipelines start with an identical module for vocabulary definition (top). Orange rounded blocks represent processes. Green rectangles represent outcomes; *pctge.* refers to the percentage of positive articles at a given tree level

million bilingually aligned Wikipedia articles for 253 language pairs, and two parallel corpora. Parallel titles can also be obtained with a tool from LTI/CMU.[8]

3 Models for article selection in Wikipedia

We tackle the automatic extraction of domain-specific comparable corpora using two alternative approaches. As far as the tools to perform standard preprocessing are at hand, both approaches are language independent and can be applied to any domain without *a priori* information. The domain is characterized by a vocabulary. The user can give an input vocabulary or allow `WikiTailor` to use the hierarchy and markup of Wikipedia to extract it automatically. Next, we describe the automatic vocabulary definition, which we use as input to both approaches, and then, we describe the approaches themselves. Figure 1 shows the pipelines schematically. Appendix 1 shows a summary description and default values of all the free parameters in the models as implemented in the `WikiTailor` toolkit.

3.1 Vocabulary definition

We extract automatically the characteristic domain vocabulary V. The input is the Wikipedia category graph G and the category c_r that better represents the desired domain (e.g., *Sport*). Our vocabulary definition and the graph exploration process depart from such node, the root category. In a first step, we select every article belonging to category c_r. The resulting set of articles, the root articles, is the seed for the in-domain vocabulary generation. If the amount of root articles R is small ($R < 10$ in our experiments), we include those articles associated with the children categories as well. In a second step, the resulting articles are concatenated into one single document and we apply the following preprocessing operations: tokenization,

[8] https://github.com/clab/wikipedia-parallel-titles.

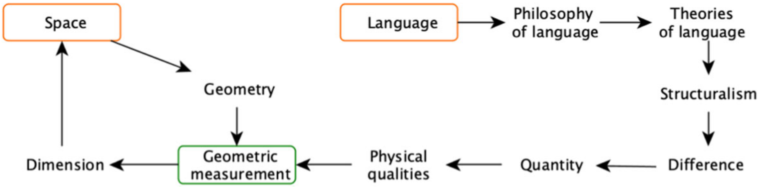


Fig. 2 Slice of the English WCG as in June 2020 departing from categories *Space* and *Language*. Both graphs meet at category *Geometric measurement* (depth 2 and 7, respectively)

stopword removal, numbers, diacritics and punctuation marks removal, and stemming [39]. Tokens shorter than four characters are discarded to reduce noise (we threshold at three for Arabic as most roots in this language are triliteral [14, p. 4]). The output consists of a list of terms ranked by term frequency. The size of this list is the free parameter that we explore in our experiments (Sect. 4).

3.2 Graph model

In this approach, we take advantage of the categories associated with Wikipedia articles. As aforementioned, even if imperfect, these categories offer important hints on the domain an article belongs to. Ideally, the categories and subcategories should compose a category tree, and one could traverse the tree to extract the related categories hanging from a specific domain (the root category). Nevertheless, the categories in the Wikipedia compose a densely connected graph G which traversal is not trivial. Figure 2 is an example of the intrinsic difficulties inherent to WCG topology (although this example comes from the English edition, others show similar phenomena). First, the paths from the unrelated categories *Space* and *Language* converge in common nodes early in the graph: *Geometric measurement*. As a result, *Geometric measurement* and all its descendants would be considered as a subcategory of both *Space* and *Language*. The topic of the root category gets diluted as we go deeper into the graph, and it can change to another topic. The 6th level departing from *Language* in this path already talks about physics. Second, G contains cycles, as observed in the sequence *Space* → *Geometry* → *Geometric measurement* → *Dimension* → *Space*.

This example evinces that one cannot consider Wikipedia's category pseudo-tree from a root category to its leaves to define a domain. Therefore, we design a strategy to walk through the graph departing from a root category to the level (depth) that most likely represents an entire knowledge domain. We tailor the Wikipedia to fit our purpose; that is, to build a well-formed tree representing a domain. Figure 1(a) depicts our graph model, which we describe below. The input consists of the domain of interest c_r, the full category graph G and the vocabulary V.

Graph article selection The module explores the category graph to find those categories which are likely to belong to the desired domain and extracts the associated articles. We perform a breadth-first search departing from node c_r. Different criteria can be considered to stop the search and prevent the exploration of the entire graph. Our stopping criterion is a heuristic inspired by the classification tree-breadth first search model by [13]. The objective is scoring the explored categories in order to assess their likelihood of actually belonging to the desired domain. We assume that a category belongs to the domain only if its title contains at least one of the words in the vocabulary. Nevertheless, many categories exist that may not include any of the words in the vocabulary. A naïve but efficient solution is to consider subsets of categories according to their depth with respect to the root and include or exclude the full subset (level). Therefore, we traverse G and score each tree level by measuring the percentage of its categories that are associated with the domain by means of containing at least one term of the vocabulary in the title. The process stops when less than $k\%$ of the categories are related to the vocabulary. In Figure 1(a), both categories in the first level fulfill the constraints, two out of three do in the second level, and three out of five do in the third one. In the fourth level, only four out of nine categories include a characteristic term in their titles. Assuming a threshold of 50%, that level in the tree is discarded and all the articles associated with the categories up to the third level compose the output.

This method has one free parameter: the percentage of categories k with an in-domain term in the title that we require to include a level in the extraction. The optimal depth for a desired domain is then determined automatically.

3.3 IR model

The authors in [38] proposed a model to retrieve Wikipedia articles associated with a domain based on a typical search engine (see Sect. 2). We implement a similar method that consists of two steps as depicted in Fig. 1(b): indexing and article selection. The input consists of the vocabulary of the domain V and the raw texts of the Wikipedia edition in the desired language.

Article indexing In an offline preliminary process, we index every Wikipedia edition and setup a search engine (right side of the bottom block in Fig. 1b). We use Apache Lucene[9] and perform a preprocessing pipeline identical to the one in the graph model.
IR article selection We query the search engine with the vocabulary V and retrieve the set of articles that presumably belong to the domain of interest.

The IR article selection method has one free parameter: the threshold on the Lucene score for the relevance of the articles.

4 In-domain collection extraction

In this section, we explore the collections obtained with the two models. We start by describing the experimental framework where they are going to be evaluated.

4.1 Framework and domains definition

We select ten Wikipedia editions that serve as archetypes for different development levels in terms of amount of articles and richness of contents: English, French, Spanish, German, Arabic, Romanian, Catalan, Basque, Greek, and Occitan. This set covers different language families, including Germanic, Romance, and Semitic. We use dumps[10] of the ten language editions from January and February 2015 and preprocess them with JWPL [58].[11] We consider

[9] Lucene is an open-source search engine: https://lucene.apache.org.

[10] https://dumps.wikimedia.org.

[11] https://dkpro.github.io/dkpro-jwpl/.

Table 1 Number of elements in the ten Wikipedia editions ranked by number of categories. Cumulative intersections are measured with respect to the languages below a given row

Edition	Articles	Categories	Ratio arts./cats.	Intersecting categories	Intersecting articles
English	4,514,317	1,206,065	3.7	–	–
French	1,487,637	303,156	4.9	141,994	933,082
Spanish	1,070,407	261,681	4.1	72,263	421,008
German	1,563,831	224,826	7.0	38,038	285,475
Arabic	331,187	122,195	2.7	19,115	87,571
Romanian	255,667	95,657	2.7	12,776	39,182
Catalan	435,817	55,099	7.9	5467	31,666
Basque	249,400	44,879	5.6	4409	19,797
Greek	100,703	30,655	3.3	3336	12,539
Occitan	90,270	15,518	5.8	2081	6811

content articles from the main namespace only, discarding redirection and disambiguation pages.[12] Table 1 shows statistics of the resulting collections.

In our work, specifying a domain is equivalent to specifying a root category for the exploration. We select automatically a set of categories that might describe the most useful and meaningful domains to analyze the performance of the models. Following [27], we look for *globally relevant concepts* for this purpose. A category is a globally relevant concept if it appears in all ten languages. Applying this constraint produces a pool of 2081 categories (cf. Table 1). We further eliminate categories starting with the same word, keeping only one of the family in any of the languages. The aim is to gather a more heterogeneous and general set, since categories that begin with the same word are usually specifications of a more general category (e.g., *Sport*, *Sport in Denmark*, *Sport in Moldova*, *Sport in New Zealand*). Categories beginning with a digit are eliminated for similar reasons. This results in a collection of 741 categories. For comparison purposes, categories used in previous research are added if not already present: *Archaeology, Linguistics, Physics, Biology*, and *Sport* [22]; *Mountaineering* [38] and *Computer Science* [5]. Observe that *Computer Science* does not exist in the Greek edition nor *Mountaineering* in the Occitan one. With these additions, we end up with 743 core domains.

4.2 Nomenclature and systems definition

From now on, WikiTailor (WT) refers to the selection method based on graphs and IR to the one based on information retrieval techniques. In the case of WT, we analyze collections extracted according to two parameters: (*i*) the percentage of categories with an in-domain vocabulary term in the title required to extract a level of the tree: we consider 50% and 60%; and (*ii*) the size of the in-domain vocabulary: we consider the top 10% of the ranked terms, and the top 100 or 500 items within the 10%. Smaller vocabularies could exist when the top-10% ranked items do not include 100 or 500 items. Table 2 shows a quick overview of the

[12] Most of such articles are labelled as such in the dumps, but some instances lack any labeling. We apply some heuristics with the aim of discarding such unlabelled, still undesired, instances. That includes the search of patterns such as `{{numberdis}}` in the title or `{{disambig}}` in the article body.

Table 2 Model settings explored. WT's settings depend on the threshold of the percentage of categories with in-domain vocabulary terms in the title (thres$_{WT}$) and the size of such vocabulary; IR's depend on the engine relevance score threshold (thres$_{IR}$) and the vocabulary size to query the engine. ⌞ indicates that the vocabulary is a subset of the top 10%

WT model	thres$_{WT}$	|vocabulary|	IR model	thres$_{IR}$	|vocabulary|
50-WTall	50%	top 10%	100-IRall	–	top 100
50-WT100	50%	⌞ top 100	100-IR10	max/10	top 100
50-WT500	50%	⌞ top 500	100-IR100	max/100	top 100
60-WTall	60%	top 10%	50-IR10	max/10	top 50
60-WT100	60%	⌞ top 100	50-IR100	max/100	top 50

setting combinations and the naming conventions. Later in the paper, we refer to subgroups of these settings with wildcards: 50-WT*, 60-WT*, *-WT100, and *-WTall.

For IR, we query the engine with the top 100 or 50 terms. The first threshold allows for a direct comparison with [38]. In their case, the characteristic vocabulary is defined as the 100 most frequent words (not terms) in an external corpus. Our IR model is clearly inspired by theirs, but we try to keep all the requirements fulfilled inside the Wikipedia itself; hence, we avoid using external corpora. In the experiments, we build the collection with all the retrieved articles (IRall), those with a relevance score higher than a hundredth of the maximum (IR100) or those with a relevance higher than a tenth of the maximum (IR10). The combined nomenclature and the usage of wildcards is equivalent to WT models. It is summarized in Table 2.

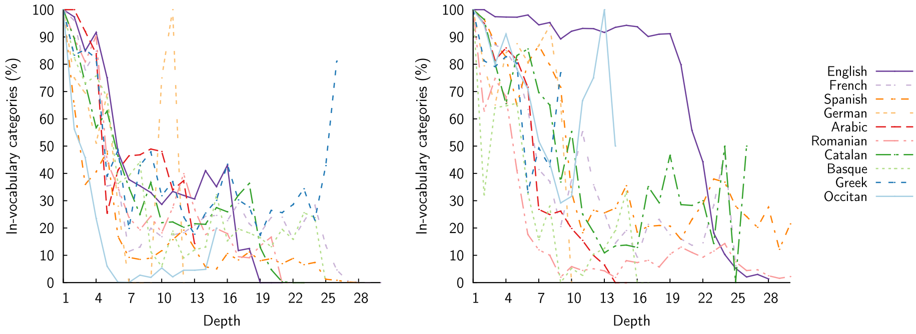
4.3 Characteristic vocabulary

The first step in both architectures is the extraction of the domain characteristic vocabulary. Following the pipeline described in Sect. 3.2, we extract the vocabularies for different language editions in the 743 categories (domains). Table 3 shows statistics on the number of articles and vocabulary sizes. As a general trend, the number of root articles, those that belong to the root category, diminishes with the size of the Wikipedia edition (except for German and Arabic). Notice that even for the largest edition (English), the mean of the number of root articles is 99, but the mode is as low as 2. Therefore, in many domains, the root articles are not enough to obtain a vocabulary large enough. This is dimmed by adding the articles in the subcategories when less than 10 articles belong to the root. In general, there is a chain relation: the larger the edition, the larger the amount of articles in the root category. This results in more terms and larger vocabularies, potentially inducing to vocabularies with a lot of noise, for large editions or for editions with many root articles, such as German. Since the quality of this vocabulary is a core factor in our methods, we explore several alternatives in our experiments. Taking the top-10% of the terms, the size of the vocabulary is completely language-dependent. Something similar happens with 500 elements, since the cut only affects major languages. For the last configuration, with a maximum of 100 elements, the size of the vocabulary is on average the same for all the languages.

We now study the distribution of this vocabulary along the graph. Recall that we consider that a category belongs to the desired domain if it has an in-vocabulary term in its title. Figure 3 depicts the evolution in the percentage of categories supposedly associated with the *Astronomy* (3a) and *Sport* domains (3b) in the ten languages. As expected, the farther

Table 3 Number of articles per category used to build domain vocabularies (mean x, standard deviation σ_x and mode m) for the ten Wikipedia editions and the 743 domains. For categories with less than 10 articles in the root, the children are considered too. The last two columns show the size of the vocabulary when the top 10% of the terms are considered

	# root articles			# (root articles +children)			Max. Vocabulary (top 10%)	
	x	σ_x	m	x	σ_x	m	x	σ_x
English	99	1332	2	533	3710	10	1154	2030
French	45	75	7	304	2710	13	755	1336
Spanish	39	145	2	141	750	14	561	720
German	193	2104	2	405	2502	10	1641	3417
Arabic	46	76	10	81	239	10	461	488
Romanian	20	39	6	56	177	12	301	409
Catalan	28	36	18	87	527	12	294	266
Basque	17	90	2	47	135	12	187	225
Greek	15	24	8	42	182	10	299	372
Occitan	8	27	1	22	80	1	102	185



(a) Domain *Astronomy* **(b)** Domain *Sport*

Fig. 3 Percentage of categories associated with two domains, according to the criteria described in Sect. 3.2 as a function of the distance to the root category

the level from the root, the lower the extent of associated categories (but also the larger the amount of elements). Peaks at deeper levels can appear due to the noisy category structure of the Wikipedia since, after departing from the original domain, the path might return to it (e.g., peak at level 13th for *Sport* in Occitan or at level 12th for *Astronomy* in German). Nevertheless, the distribution is rough and, at the lowest levels, the small number of articles can lead to artificial canyons in the curves (e.g., canyon at the 2nd level for *Sport* in Basque). This effect is domain- and language-dependent. We deal with more than 7,000 domains (743 domains × 10 languages). Hence, on average the effect is not important and all the process is done fully automatically. However, to obtain a corpus in a concrete language and domain, a visual inspection of the shape of this curve helps to determine the halting point.

Table 4 Selected depth threshold per category (mean x, standard deviation σ_x and mode m) for the ten Wikipedia editions and the 743 domains

	50-WT100			50-WT500			50-WTall			60-WT100			60-WTall		
	x	σ_x	m	x	σ_x	m	x	σ_x	m	x	σ_x	m	x	σ_x	m
English	5.9	2.8	5	9.6	7.3	8	12.4	11.1	8	5.2	2.4	7	10.5	9.6	[7, 8]
French	4.3	1.9	5	5.1	2.3	5	5.7	3.6	5	3.8	1.9	5	5.2	3.2	5
Spanish	4.4	2.1	2	6.0	3.7	2	6.9	5.8	2	3.8	1.9	2	5.8	4.4	[2, 6]
German	3.4	1.9	2	3.8	2.1	2	4.0	2.2	2	3.1	1.8	2	3.8	2.2	2
Arabic	3.6	2.3	1	4.7	3.7	1	6.1	4.6	5	2.9	2.0	1	5.2	3.4	5
Romanian	3.4	1.8	2	3.8	2.1	2	3.8	2.1	2	3.2	1.6	2	3.6	2.0	2
Catalan	3.3	1.9	2	3.8	2.2	2	3.8	2.3	2	2.9	1.8	2	3.4	2.1	2
Basque	3.1	1.5	2	3.3	1.7	2	3.3	1.7	2	2.8	1.4	2	3.1	1.6	2
Greek	3.0	1.6	2	3.3	1.8	2	3.3	1.9	2	2.8	1.5	2	3.1	1.8	2
Occitan	2.4	1.3	2	2.5	1.4	2	2.5	1.4	2	2.2	1.2	2	2.4	1.3	2

4.4 Collections characteristics

`WikiTailor` determines automatically the depth up to which it should extract articles according to the percentage of in-vocabulary categories. This is a crucial point for the extraction: different percentages lead to different stopping points and collection sizes. Looking at the numbers in Table 4, the **threshold depth** seems to be directly proportional to the size of the characteristic vocabulary and the number of categories. In general, the more categories in a Wikipedia edition, the more levels are used to describe a root category. These two features are more important than the alternatives of taking levels with a 50% or a 60% of positives. For a given language, the most relevant feature is the size of the vocabulary, specially for small editions: smaller vocabularies imply smaller depths. For Romanian, Catalan, Basque and Greek, systems with 50% of positives select a mean boundary depth of 3 for WT100 and 4 for WTall. The change is less significant for the systems with 60%. In the large editions, the change is striking in both cases. In English, systems with 50% of positives select a mean threshold depth of 6 for WT100 and 12 for WTall (5 and 11 for the 60% systems). Hence, for the editions with more articles, we extract the articles from a larger subtree, favoring the extraction of huge in-domain corpora for English and more modest ones for the other languages. As before, Arabic and German seem to be out of place. If we rank the editions according to the number of categories, Arabic has a higher-than-expected mean depth per domain. German has it lower. All differences among languages are reduced for small and similar vocabularies (*-WT100).

The top rows of Table 5 show the **size of the collections** extracted with WT. The size for every system and language is a direct consequence of the aforementioned. Except for Arabic and German, the larger the edition, the larger the extracted collection of in-domain articles, but, for small vocabularies, the differences are less extreme. The loss in number of articles in English for small vocabularies respect to *-WTall is remarkable (from 1 M in 50-WTall to 50 k in 50-WT100). This is not the case for German (5 k vs 3 k) although its initial vocabulary for 50-WTall was even larger than the English one.

The bottom rows of Table 5 describe the in-domain corpora extracted with the IR model. In general, IR retrieves larger collections than WT, up to the point that for queries with 100 terms and without any threshold for the relevance score (IRall) the extracted corpus approaches

Table 5 Mean N and standard deviation σ_N of the number of articles per domain for WT (top) and IR models (bottom). We show five systems with different values for the two free parameters in both cases (cf. Section 3). Left-most numbers indicate the ranking of the edition in number of articles (cf. Table 1)

		50-WT100		50-WT500		50-WTall		60-WT100		60-WTall	
		N	σ_N	N	σ_N	N	σ_N	N	σ_N	N	σ_N
1	English	50,514	121,881	513,615	1,170,041	1,008,340	1,780,484	27,903	59,838	734,168	1,544,054
3	French	8278	26,483	18,518	79,134	40,207	182,940	5869	21,342	29,717	147,592
4	Spanish	6638	17,050	34,556	97,268	63,450	166,908	4463	14,257	36,445	120,180
2	German	2752	9573	4131	16,658	5150	19,860	2199	8948	4671	19,184
6	Arabic	2999	9,546	20,441	58,483	36,969	87,961	1541	6354	20,456	60,213
7	Romanian	1398	8683	2078	11,875	2396	13,649	796	3590	1766	11,207
5	Catalan	1140	4693	2041	9891	2319	10,991	686	2317	1601	7889
8	Basque	440	1654	907	4673	1130	6638	320	1215	748	4207
9	Greek	390	2008	777	4218	833	4649	298	1407	679	4097
10	Occitan	104	598	247	2480	308	3051	68	374	114	774
		100-IR10		100-IR100		100-IRall		50-IR10		50-IR100	
		N	σ_N	N	σ_N	N	σ_N	N	σ_N	N	σ_N
1	English	64,239	73,248	1,119,637	482,339	3,947,077	221,129	52,030	69,135	976,547	488,533
3	French	18,158	17,871	331,936	145,261	1,235,344	88,852	15,258	18,115	308,093	152,543
4	Spanish	21,490	19,605	314,612	105,193	958,399	50,521	17,791	20,454	283,654	113,708
2	German	12,887	18,876	378,195	218,271	1,434,164	70,250	12,843	21,628	344,392	225,517
6	Arabic	4622	4082	61,882	25,026	274,589	36,760	4188	4,020	58,284	24,957
7	Romanian	1750	1839	33,018	15,213	162,608	32,409	1567	1882	32,199	16,364
5	Catalan	5959	5058	99,703	38,195	370,917	35,443	5601	5940	97,653	39,756
8	Basque	1819	2732	33,173	21,743	139,099	33,196	1741	2889	33,421	23,174
9	Greek	2547	2378	40,074	14,550	97,417	46,153	2639	3349	41,396	16,108
10	Occitan	419	2040	6397	7052	42,554	21,270	487	2331	6617	7125

the full Wikipedia. The number of articles extracted by the IR models is proportional to the size of the collection, not to the number of categories, as it happens with WT and small vocabularies. As expected, queries with less items retrieve smaller collections (50-IR* vs 100-IR*). Some exceptions appear for Basque and Greek. This occurs when one does not look at the collection with all the hits (IRall) but at those recovering a percentage of the maximum score. Since the maximum score changes when using 100 and 50 query terms, the same can happen for the number of elements.

WT and IR build very different corpora, specially in terms of **content**. WT collections are smaller, but they are not subsets of the IR ones (except in the cases in which IRall is the reference, the system that selects almost the whole Wikipedia regardless of the domain). For instance, 50-WT100 and 100-IR10 have similar dimensions. Still, only between 20–60% of the WT articles and a 5–15% of the IR ones appear in the intersection between the extracted collections. The common articles cover a larger percentage of the WT collections because their size is smaller. The ranges in the previous figures describe the behavior for the different languages. Large editions have a lower percentage of common articles (for example 23% and 56% for WT in English and Greek and 8% and 4% for IR in the same languages).[13]

It is worth noting that these results correspond to the monolingual scenario. A multilingual comparable corpus is just the set of collections of the same domain for each language. We can increase the degree of comparability [22, 50] by selecting a subset of equivalent articles in a straightforward way thanks to Wikipedia's inter-language links. Once the monolingual corpora have been retrieved, the union or intersection of their linked articles constitutes the final domain-specific comparable corpus.

4.5 Comparison against similar systems

Gamallo Otero and González López [21, 22] obtained comparable corpora in Spanish, English and Portuguese in the *Archaeology, Linguistics, Physics, Biology*, and *Sport* domains based also on Wikipedia's categorization. The comparison against our model is difficult because the Wikipedia editions considered differ by six years, doubling their size during this period. Besides, they report the size of their comparable corpora in MB and not in number of articles. The single comparison we can do is that for the comparable corpus obtained for *Archaeology* in English and Spanish. Their most flexible (tight) method retrieves 1,120 (34) articles in English and 462 (34) in Spanish. Our most restrictive 60-WT100 configuration reaches depth 5 and retrieves 65,343 articles for English and depth 2 with 553 articles for Spanish. The conservative 50-WT100 retrieves 236,951 articles in English (depth 6) and 17,335 in Spanish (depth 5). Of course, the accuracy of CorpusPedia is much higher, but for some tasks the size of the corpus would not be enough. Notice that we are talking about the size of the collections, not about their quality.

The authors of [38] used a very similar method to IR to extract parallel articles in the *Alpine* domain for German and French. We can compare their results with the ones we have for *Mountaineering* with our IR model but, again, the Wikipedia editions differ. They index only aligned documents according to the inter-language links, since they aim at extracting parallel sentences which can be assumed to be mostly found in aligned documents. They retrieve 40,000 parallel articles, whereas our conservative 100-IR100 retrieves 225,422 French and 305,200 German articles. We can extract the subset of parallel articles from this comparable corpus via the intersection or the union of the articles. For the intersection, we use the articles identified as in-domain simultaneously in German and French. For the union, we expand the

[13] Tables with the percentages broken down by language and model are provided as supplementary material.

set to include all article pairs if at least one of them has been identified as in-domain in either language. Using the intersection, we obtain a high precision set with 55,551 articles and with the union we gather a high recall corpus with 205,913 articles.

4.6 Manual evaluation

We are interested in determining whether the documents in a collection belong to a particular domain or not. For this manual study, we select two representative systems: 50-WT100 and 100-IR10 and judge manually their articles in three domains in all ten languages: *Astronomy*, *Software*, and *Sport*. The evaluation set for each language, domain, and system consists of 200 articles: 100 exclusive to each system and 100 in common to both. The articles are extracted evenly in its subset. In three cases, the number of articles in the collection is smaller than 200 and so is the evaluation set (see Table 6). We manually annotate the 8,600 articles with three assessments each using the Figure Eight[14] crowdsourcing platform. Appendix 1 includes the experiment set up and instructions for the Turkers.

Table 6 shows the manually judged precision results. We calculate the precision of the extracted collections under two circumstances: (*i*) **hard precision** when there is full agreement in assigning a domain among the three annotators and (*ii*) **soft precision** when an article is assigned to a domain by two out of three. For the three domains, the quality of the WT extractions is much better than those with IR. Even in the hard-precision setting, the mean value is 0.74±0.14 for WT and 0.43±0.12 for IR. Values per domain are close to them. The average values for soft precision go up to 0.84±0.13 for WT and 0.50±0.14 for IR. Focusing on the language factor, the IR system does specially well for German, suggesting a higher vocabulary quality. This is an indication that the quality of the characteristic vocabulary is less important in WT than in IR: WT averages all the categories in a level before extracting it, dimming the negative impact of a noisy vocabulary. On the other hand, WT's weakest performance comes with Arabic, with a mean soft precision over domains of 0.57±0.11. Arabic collections are built after considering a low depth (3.6±2.3 with a mode as low as 1; cf. Table 4). Nevertheless, the three domains evaluated are built upon a higher depth (5 for *Astronomy*, 8 for *Software*, and 6 for *Sport*) meaning that perhaps too many articles are extracted increasing the coverage but damaging the precision. The outcome is still better than for its IR counterpart.

The difference between the WT and the IR systems becomes more evident when looking into the distribution of their resulting collections. As said before, we have built the subsets to evaluate by assuring that half of the articles in a collection are common in both systems and the other half is exclusive to each of them. That allows us not only to save in manual assessments, but also to have a clear idea of the distribution of the articles in a collection. The block "100-element subset" in Table 6 shows the results. As expected, the articles that are common to both systems ($\cap_{\text{only}}$) are those with the highest precision (on average 0.79±0.15 for hard and 0.89±0.15 for soft). The quality of the collections extracted only by the WT system (WT_{only}) is very close in quality with an average of 0.70±0.17 for hard precision and 0.80±0.17 for soft precision. The precision is very low for articles only retrieved by the 100-IR10 system (mean of 0.11±0.16 for hard and 0.16±0.20 for soft). The only exception is again German, where the IR_{only} subcollection has a hard precision of 0.50±0.24 and a soft precision of 0.61±0.19.

The last column of Table 6 shows the Fleiss' kappa (κ_{Fleiss}) interannotator agreement [19]. Turkers agreed the most when discriminating between *Sport* and *other*, with an average

[14] https://www.figure-eight.com/.

Table 6 Results of the manual evaluation. Set_{WT} and Set_{IR} show the number of articles selected for manual assessment. “Complete set” shows the precision obtained under the hard and soft criteria for the 50-WT100 (WT) and 100-IR10 (IR) systems. “100-element subset” shows the distribution of the sets (see text). κ_{Fleiss} shows the inter-annotator agreement

			Complete set				100-element subset						
	Set_{WT}	Set_{IR}	WT		IR		WT_{only}		IR_{only}		$\cap_{\text{only}}$		κ_{Fleiss}
			hard	soft	hard	soft	hard	soft	hard	soft	hard	soft	
ASTRONOMY													
English	200	200	0.67	0.82	0.37	0.44	0.64	0.84	0.05	0.07	0.70	0.81	0.789
French	200	200	0.71	0.81	0.40	0.45	0.69	0.80	0.07	0.07	0.74	0.83	0.828
Spanish	200	200	0.83	0.94	0.45	0.55	0.83	0.87	0.07	0.10	0.83	1.00	0.834
German	200	200	0.77	0.92	0.76	0.87	0.72	0.88	0.71	0.79	0.82	0.96	0.524
Arabic	200	200	0.59	0.65	0.41	0.45	0.37	0.46	0.02	0.05	0.81	0.85	0.816
Romanian	200	200	0.76	0.82	0.41	0.47	0.71	0.72	0.01	0.02	0.81	0.93	0.883
Catalan	200	200	0.73	0.85	0.46	0.51	0.66	0.79	0.11	0.11	0.81	0.92	0.786
Basque	200	200	0.95	0.99	0.48	0.52	0.96	0.98	0.03	0.05	0.94	1.00	0.929
Greek	200	200	0.70	0.74	0.44	0.48	0.52	0.57	0.00	0.04	0.88	0.92	0.893
Occitan	132	139	0.66	0.71	0.35	0.40	0.59	0.65	0.16	0.21	0.85	0.90	0.845
SOFTWARE													
English	200	200	0.73	0.92	0.42	0.56	0.69	0.91	0.07	0.19	0.78	0.94	0.657
French	200	200	0.77	0.91	0.42	0.54	0.82	0.92	0.12	0.18	0.72	0.91	0.735
Spanish	200	200	0.80	0.93	0.48	0.56	0.75	0.93	0.12	0.17	0.85	0.95	0.756
German	200	200	0.72	0.84	0.52	0.65	0.63	0.80	0.24	0.41	0.81	0.90	0.609
Arabic	200	200	0.41	0.61	0.30	0.47	0.26	0.42	0.04	0.15	0.56	0.79	0.614

Table 6 continued

			Complete set				100-element subset						
	Set_{WT}	Set_{IR}	WT hard	soft	IR hard	soft	WT_{only} hard	soft	IR_{only} hard	soft	$\cap_{\text{only}}$ hard	soft	κ_{Fleiss}
SOFTWARE (continued)													
Romanian	200	200	0.84	0.97	0.49	0.58	0.84	0.98	0.13	0.19	0.85	0.97	0.788
Catalan	200	200	0.80	0.93	0.50	0.60	0.77	0.93	0.18	0.26	0.83	0.94	0.724
Basque	200	200	0.86	0.96	0.48	0.53	0.83	0.95	0.08	0.09	0.89	0.98	0.852
Greek	200	200	0.87	0.96	0.51	0.56	0.86	0.95	0.13	0.15	0.89	0.97	0.928
Occitan	17	73	0.72	0.89	0.30	0.58	0.82	0.91	0.27	0.55	0.67	1.00	0.424
SPORTS													
English	200	200	0.70	0.79	0.44	0.48	0.55	0.64	0.01	0.01	0.87	0.95	0.897
French	200	200	0.92	0.93	0.50	0.50	0.86	0.88	0.02	0.02	0.98	0.98	0.947
Spanish	200	200	0.86	0.90	0.49	0.50	0.77	0.83	0.03	0.03	0.95	0.97	0.948
German	200	200	0.55	0.75	0.56	0.66	0.66	0.82	0.55	0.64	0.47	0.51	0.710
Arabic	200	200	0.37	0.44	0.13	0.16	0.51	0.58	0.02	0.02	0.24	0.31	0.877
Romanian	200	200	0.73	0.78	0.37	0.41	0.72	0.74	0.01	0.01	0.74	0.82	0.920
Catalan	200	200	0.60	0.68	0.33	0.38	0.56	0.61	0.02	0.02	0.65	0.75	0.876
Basque	29	111	0.87	0.97	0.10	0.12	0.94	1.00	0.02	0.02	0.82	1.00	0.943
Greek	200	200	0.91	0.96	0.44	0.48	0.94	0.96	0.00	0.01	0.89	0.96	0.841
Occitan	200	200	0.74	0.77	0.56	0.57	0.50	0.54	0.14	0.14	0.99	1.00	0.946

$\kappa = 0.88 \pm 0.07$. The lowest agreements occurred in the *Software* domain: $\kappa = 0.74 \pm 0.11$. *Astronomy* lies in the middle with 0.81 ± 0.12. Regarding the language dimension, annotators of Basque agreed the most, with $\kappa = 0.91 \pm 0.05$. Instances in German were the least agreed upon, with $\kappa = 0.61 \pm 0.09$. Individually, annotators of Spanish instances *Sports* vs *other* obtained the highest agreement: 0.95. The lowest agreement was obtained for *Astronomy* vs *other* in German, 0.52. Notice that in most cases, 28 out of 30, we obtain either *substantial agreement* ($0.61 < \kappa < 0.80$) or *almost perfect agreement* ($0.81 < \kappa < 1.00$) as defined in [31]. We can conclude that 50-WT100 is significantly better than 100-IR10. However, a manual evaluation is always expensive and one would like to quantify automatically the adequateness of a collection with respect to the desired domain. Next section introduces the concept of domainness to address the issue.

5 *Domainness* characterization

We are interested in determining automatically whether the documents in a collection belong to a particular domain or not. Still, describing corpora is a difficult and subjective task and the answer should not be binary, but a continuous score, especially if it is quantified automatically. We define domainness as the degree of cohesion and representativity of a corpus with respect to a domain:

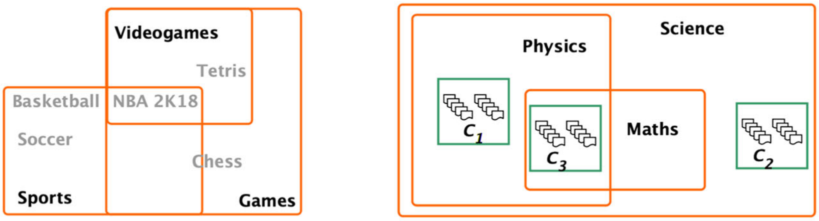
$$\text{domainness} = \text{representativity} + \text{cohesion}$$

The idea behind the domainness concept builds on the intuition that a collection should be heterogeneous but cohesive at the same time. For illustrative purposes, Fig. 4a shows three domains and five Wikipedia articles within them. Article *Basketball* clearly belongs to domain *Sport*, whereas *Tetris* clearly does not. Articles such as *NBA 2K18* lie within all *Sport*, *Games* and *Videogames* domains. Yet the membership of *NBA 2K18* in the *Sport* domain is subjective, unless a more detailed description of the domain is given. A collection with these three documents is less representative of *Sport* than one including articles *Basketball*, *Soccer* and *Chess*, which are more **cohesive**. To what extent remains subjective; we need a measure to quantify the difference.

Figure 4b shows another example to illustrate the concept of **representativity** within a collection. Whereas collections C_1 and C_3 correspond to the *Physics* domain, C_1 should receive a higher domainness score because articles seem to be purely about physics (C_3 contains articles in the intersection of physics and math). When measuring the domainness of the collections with respect to the *Science* domain, C_3 should have a higher value because it has more diversity, i.e., it holds a higher representativity of the domain. In this scenario, one cannot say which of C_2 or C_3 should have a higher domainness score for *Science*.

To the best of our knowledge, no specific measures exist to quantify this concept. Although there is no predefined scale to quantify domainness either, we intend to measure if a corpus represents better a domain than another one, and how or if it degrades when enlarged. To produce an affordable evaluation framework, we define four families of automatic metrics inspired by the work of [29] on corpus analysis and the work of [35] on topic coherence. The first three families measure the representativity of the corpus and characterize a domain on the basis of its characteristic vocabulary. Quite differently, the fourth family measures the cohesion of the collection without the requirement of characterizing the domain.

Family 1: Density of terms We begin with the assumption that the higher the density of the characteristic vocabulary in a corpus, the better it describes the domain. Obtaining this



(a) Intersecting domains (orange boxes) and example articles within them (in gray).

(b) Collection C_1 has the highest domainness for *Physics*; C_3 and C_2 have higher domainness for *Science* since they have a major representativity.

Fig. 4 Representativity and cohesion as a measure of domainness

vocabulary is straightforward when using the Wikipedia as a corpus. Since root articles belong to the domain by definition, the characteristic vocabulary can be obtained as the most frequent terms in this subcorpus. The *density* of these terms should be a measure of the representativity of the collections. We propose two densities based on two term frequency estimations [45]. The first one is the term frequency of all in-domain terms w_i in the collection, $c_{\text{terms}}=\sum_{w_i}\text{counts}(w_i)$, normalized by the number of articles, N:

$$C_{\text{terms}}/N \equiv \frac{1}{N}\sum_{\text{art}} c_{\text{terms}}. \tag{1}$$

The second one is the augmented frequency of in-domain terms for each article normalized by the number of articles:

$$\hat{c}_{\text{terms}} = \frac{1}{N}\sum_{\text{art}}\left(K + (1-K)\frac{c_{\text{terms}}}{c_{\text{max}}}\right), \tag{2}$$

where c_{max} are the counts for the most frequent term in each document and the optimum value of K is 0 in our experiments.

Family 2: Mutual information The quality of a corpus in terms of domainness is somehow related to the evaluation of topic models. In the first case, we have a collection of texts and we want to evaluate how well they describe a domain that might be characterized or not by a set of keywords. In the second case, we are given a set of keywords and we want to evaluate how well they describe the topic (domain) of a collection. The authors in [35] introduced the concept of *coherence* of a topic as the interpretability of its keywords. They measure it with the average or median of the pointwise mutual information (PMI) among the topic keywords. Subsequent works use NPMI [7], a normalized version of PMI:

$$\text{PMI}(w_i, w_j) = log_2\frac{p(w_i, w_j) + \epsilon}{p(w_i)\, p(w_j) + \epsilon}, \tag{3}$$

$$\text{NPMI}(w_i, w_j) = \frac{\text{PMI}(w_i, w_j)}{-log_2(p(w_i, w_j) + \epsilon)}, \tag{4}$$

where w_i and w_j are the keywords describing a topic—the terms in the characteristic vocabulary in our case—ϵ is a smoothing constant, and p stands for frequentist probability. For topic modelling, the median of the pairs showed better correlation with human judgments than its mean because it is less sensitive to outliers [35].

We apply the two measures and two variants to evaluate domainness; assuming that the vocabulary we use perfectly describes the domain and the loss in the value of (N)PMI gives information about the background collection. We expect in-domain collections to show a high in-domain terms density—$p(w_i)$ and $p(w_j)$ values higher than in general collections—but we still expect co-occurrences of terms to be representative. Computationally, the main difference with the original usage is how to estimate term co-occurrence frequencies to compute probabilities. In topic modelling, co-occurrences are sampled from the full collection or from an external source, such as the Wikipedia or Google n-grams, with a sliding window of m words. Here, we always use the full in-domain collection and consider as window an entire article of the domain: (N)PMI$_{\text{art}}$. With this definition, the window has a variable length. To study if this difference is relevant, we define a second variant (N)PMI$_{\text{col}}$ where we estimate a probability as the sum of probabilities in all the articles of the collection instead of simply the counts per article as in the original version:

$$p(w_i)_{\text{art}} = \frac{\sum_{\text{art}} \text{counts}(w_i)}{\sum_{\text{art}} \text{terms}} \quad \text{and} \quad p(w_i)_{\text{col}} = \frac{1}{N} \sum_{\text{art}} \frac{\text{counts}(w_i)}{\text{terms}}. \tag{5}$$

Family 3: Correlations The authors in [29] quantifies the similarity among corpora by measuring frequencies of words and cross-entropies. We adapt his best measure to fit our problem, the Spearman correlation, and add Kendall's τ correlation for a better generalization. Spearman ρ (and Kendall's τ) is a nonparametric rank correlation. It measures the difference in rank order between two distributions:

$$\rho = 1 - \frac{6 \sum pd_i^2}{n(n^2-1)}, \tag{6}$$

where pd are the pairwise distances of the ranks of the terms w_i and w_j, and n is the number of terms. For Kendall, we have:

$$\tau = \frac{c-d}{\sqrt{n(n-1)/2 - T}\sqrt{n(n-1)/2 - U}}, \tag{7}$$

where c is the number of concordant pairs, d is the number of discordant pairs, and

$$T = \sum_t t(t-1)/2 \text{ and } U = \sum_u u(u-1)/2, \tag{8}$$

where t is the number of times the terms w_i are tied and u is the number of times the terms w_j are tied.

In our case, we measure the difference in rank order of n terms in two corpora: an extracted collection of articles of a given domain and the subset of its root articles. Terms are defined as before; since the important feature of a term is its rank and not its absolute frequency, this measure can be used for corpora of varying size.

To compute the correlation, one needs to find the n most frequent common terms. These are obtained as the union of the first m terms for every corpus. The terms that the other corpus lack have frequency zero and are therefore ranked at the bottom of the other corpus' list. Some heuristics are considered to build the vectors: (*i*) At most 1000 terms from the top 10% (if available) for every collection are used, therefore the maximum number of common elements is 2000; (*ii*) terms with frequency 1 are not considered; and (*iii*) correlations are not estimated with less than 5 points.

Both Spearman and Kendall correlations measure monotonicity relationships. Although we checked that in most cases the two statistics lead to the same conclusions, Kendall's τ is the representative of this family since it has shown to be more robust, more appropriate for small samples and, given its definition, to deal better with ties and outliers [12].

Family 4: Cohesion In this case, our objective is assessing the distance between the articles that belong to a given domain. The lower the distance between such articles, the more cohesive they are, and the more likely that they actually belong to the domain; i.e., the better the model works. In order to come out with a single number to compare across different models, we compute the average distance between all the article pairs in the domain. Considering standard vector-space models to represent the texts could result in measures sensitive to length and vocabulary differences between the pairs of articles. Article embeddings obtained as document embeddings simply by using doc2vec [32] could solve this issue, but the quality would still depend on the language because low-resourced languages have less data to estimate the embeddings. Since we focus on multilinguality, we opt for using ESA, a high-dimensional concept-based representation.

ESA represents texts—regardless of their lengths—onto a high-dimensional concept-based space. The space is built on top of the term–document matrix **D** generated from a large collection D of documents using tf-idf weighting. The representation of a text is then built by comparing it against **D**, resulting in a $|D|$-dimensional vector. For efficiency reasons, the average distance is computed with respect to the center of the collection as

$$d_{\mathrm{ESA}} = \frac{1}{N} \sum_{a} dist_{\theta}(a_{ESA}, c_{ESA}), \tag{9}$$

where a_{ESA} is the vector representing article a, c_{ESA} is the centroid of all the vectors in the corpus, and $dist_{\theta}$ refers to the angular distance:

$$dist_{\theta} = \arccos\left(\frac{a_{ESA}.c_{ESA}}{\| a_{ESA} \| \| c_{ESA} \|}\right). \tag{10}$$

6 *Domainness* evaluation

Now, we inspect the numbers obtained for the different metrics when analyzing the collections extracted by the WP and IR models in all languages and domains. Figure 5 summarizes the results with some representative measures from the four families of metrics.[15] We plot the mean and standard deviation of six measures: C_{terms}/N, $\hat{c}_{\mathrm{terms}}$, $\mathrm{PMI}_{\mathrm{art}}$, $\mathrm{PMI}_{\mathrm{col}}$, τ, and d_{ESA}, for the ten systems analyzed. For comparison purposes, we also chose a representative model of every family (50-WT100 and 100-IR10) and compare it against a subcollection of the other family gathered to have the same size. Although we do not include the corresponding figures, the outcomes are also discussed. For the representativity measures (Families 1, 2 and 3), the size of the characteristic vocabulary used in the experiments is 100 terms, i.e., 5,049 term pairs. In all cases, the collections on which probabilities are estimated are preprocessed as explained in Sect. 3.2 so that the format of the articles matches the terms.

Family 1 By design, IR systems have the largest number of in-domain terms. The density is expected to be higher in the smallest *-IR10 collections because they contain the top ranked articles according to these terms. Also by definition, a high density of terms exists in the root articles of the WT systems, but there is no expectation for a high number of in-domain terms in the rest. The output of $\hat{c}_{\mathrm{terms}}$ and especially of C_{terms}/N reflects this (cf. Fig. 5 top-left plot). Differences between WT systems do not seem significant under these metrics.

[15] We include as supplementary material the corresponding tables.

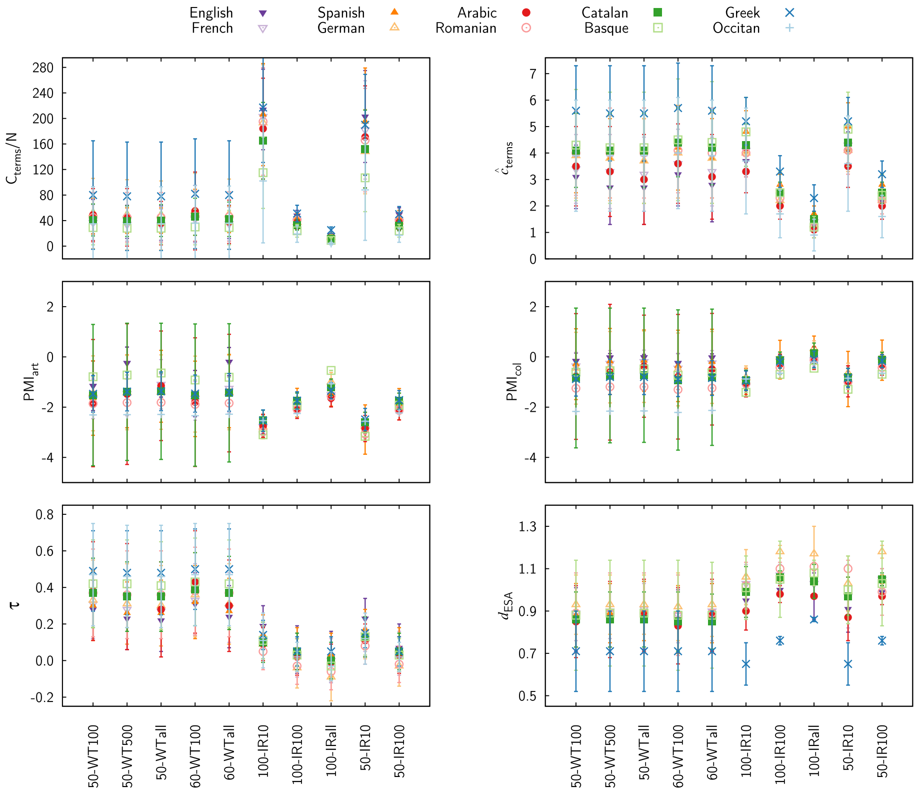


Fig. 5 Automatic evaluation of the in-domain collections for the systems and languages under study with six representative measures of the four families introduced in Sect. 5. Points represent the arithmetic mean over the 743 selected domains

In general, differences appear in large editions, where the vocabulary size varies notably across systems. The best WT system is 60-WT100, the most restrictive and the one with less articles per collection: a mean of $C_{\text{terms}}/N = 49.7$ and $\hat{c}_{\text{terms}} = 4.1$ across languages. However, 60-WTall has a higher density of in-domain terms than any of the 50-* systems for some editions (those with less categories) even if the resulting corpora are larger.

According to C_{terms}/N, IR systems with the smallest collections (*-IR10) are clearly the best ones, as expected from its definition. The normalization in $\hat{c}_{\text{terms}}$ smooths the effect and makes systems closer to each other. Since IR collections grow significantly after allowing for lower relevance scores, there are many differences between IR models. According to these metrics, *-IR10 systems have better *quality* than any WT model, especially for large editions, with the additional benefit that they gather larger collections. This effect is more pronounced when comparing equal-size collections, but disappears for the less constrained configurations where WT models are better. Regarding language, both models perform at their best in Greek. There is no clear trend for the other editions, although English and Arabic perform poorly in contrast with the others. This is one of the differences when evaluating with the correlation family of metrics (Family 3). In this case, English, Greek and Spanish are the languages with the best results. This is a first indication that both metrics are not equally valid for assessing the quality of the extractions.

Family 2 Contrary to in-domain terms, there is no requirement on the number of term co-occurrences when building IR or WT systems. The plots in the middle row of Figure 5 show the mean and standard deviation of PMI_{art} and PMI_{col}. One would expect positive PMIs for related terms, meaning that they occur more frequently together than if they were independent in a general collection, but we obtain negative values for most collections. The reason is the high density of in-domain terms in all the documents, which causes co-occurrences to have comparatively less weight than in general collections.

Since we want to indirectly evaluate the collection and not the terms, we just compare the values of the different models. Within a family of systems, WT or IR, the scores completely depend on the size of the collection: the larger the collection, the better the evaluation. WT systems are better than IR systems even if IR collections tend to be larger. For instance PMI_{art}=-1.1±1.0 for the 50-WT100 English collection, with a mean of 50,514 documents per domain, and PMI_{art}=-2.8±0.3 for 100-IR10 with a mean of 64,239 documents per domain. The values of PMI_{col} for these collections are -0.2±0.3 and -1.2±0.4. We observe the same trends with PMI_{art} and PMI_{col}, but the scores with PMI_{col} tend to be higher. Differences among models turn smaller in terms of normalized PMIs, but the main conclusions hold.

When looking at differences across languages, the scores are almost independent of the language for IR systems, whereas English collections are the best ones for WT systems and the Romanian and Occitan the worst ones. Besides, Romanian, Basque and Occitan have large deviations, especially in WT systems. In IR systems, these languages have the smallest collections, but this is not the case for WT. The uncertainties for these languages, which range from ±4 to ±8, are not shown in Fig. 5 for clarity.

Family 3 As observed in the bottom-left plot of Fig. 5, correlation measures show a clear preference for the WT model. Kendall's τ lies in the range $[0.2, 0.5]$ for WT and in $[-0.1, 0.2]$ for IR systems. Results are equivalent with Spearman's ρ although with a higher score: within $[0.3, 0.6]$ and $[-0.1, 0.3]$, respectively. For different variations of a model, the results are consistent with those seen with the measures related to the density of terms: smaller and more constrained collections are always evaluated better. However, the standard deviation is too big to make statistically significant statements when comparing models within one family. In general, the *quality* increases for Wikipedia editions that have less categories for WT systems; whereas there is no specific trend for IR systems. Large editions correlate less because their domains have more articles; when only domains with more than 100 articles are considered, correlations diminish for those languages where this is important (e.g., Occitan, Greek, or Basque) and the scores per language become more homogeneous. When we compare IR and WT collections up to an equal size, we confirm that WT models are better than the IR ones according to ρ and τ and, the smaller the edition, the more evident the difference becomes.

Family 4 Following the original ESA proposal and in consistency with this work, we use the Wikipedia as our reference text collection D for the cohesion-oriented metric. The size of D for each language is 12,539, as this is the size of the intersection among the top nine Wikipedia language editions. The authors in [23] showed the convergence of the method with 10,000 articles approximately. Hence, we discard the Occitan edition because it would significantly reduce the size of D.

Similar trends seen with the previous metrics regarding quality can be observed with d_{ESA}, even if its nature is different. In this case, lower values imply collections with a higher cohesion, irrespective of the domain they belong to. The results are shown in the bottom-right plot of Fig. 5. Since WT collections include the root articles of the desired domain and IR systems retrieve only articles that contain the vocabulary of the domain, we can assume that a large cohesion implies a large domainness. As it happens with ρ and τ, d_{ESA} clearly peaks WT models ($d_{\text{ESA}} \approx 0.85$) over IR ones ($d_{\text{ESA}} \approx 1.00$). The best (worst) collections are obtained for Greek (German). Again, mean averages do not allow to establish preferences among the different configurations within a family of models in a statistically significant way, but models with the smallest set of terms (*-IR10 and *-WT100) are preferred; i.e., more constrained collections have a larger cohesion.

All the metrics differentiate clearly the quality of WT and IR systems, but only show trends within models in a family. In general, the most constrained configuration per family (60-WT100 and 50-IR10) obtains the most in-domain collection. Still, the difference is often minimal with respect to another configuration which, on the other hand, might have retrieved many more articles. We are comparing 7,430 collections for 10 different models. In practice, one would deal with a few. In that case, it might be more fruitful to decide which is the most convenient collection according to the scores, to the size, and to the domain representativity requirements. Notice also that the density metrics (Families 1 and 2) behave differently to correlation (Family 3) and cohesion (Family 4) when dealing with the most constrained collections.

The human judgments from Sect. 4.6 allow to estimate the quality of the automatic evaluation metrics. We calculate the Pearson correlation r_P between the crowdsourced precisions and the automatic scores on the same subcollections, considering 200 articles per system and language in three domains (settings in Sect. 4.6).

A visual inspection of the data is a good first clue to understand the behavior of the metrics. Figure 6 shows the relation against soft precision for six metrics: C_{terms}/N, $\hat{c}_{\text{terms}}$, PMI_{col}, τ, d_{ESA}, and the full domainness measure *Dom*; see Eq. (11). In all cases, the graphical counterpart of Table 6 (e.g., points corresponding to the 50-WT100 system; green bullets) is located toward higher precision values than those for the 100-IR10 system (orange diamonds). We plot 60 points per figure: two systems × ten languages × three domains. The exceptions are $d_{\text{ESA}}$and *Dom*, for which only nine languages × 3 domains are shown (we discard those collections with less than 200 articles for the correlation estimation (*Astronomy* and *Software* for Occitan and *Sport* for Basque; cf. Table 6).

Family 1 Counterintuitively, the metric with the highest and negative correlation is the density of terms C_{terms}/N with $r_P = -0.716$. The high value is just an artifact given by the different composition of the WT and IR collections. By construction, the IR system retrieves articles with lots of terms, whereas the dependence for WT models is lower. The quality of WT is better, so there is a clear anticorrelation between the density of terms and the precision. If we look independently within WT or IR instances (i.e., green or orange points alone), we obtain worse correlation values: $r_P = -0.18$ for WT and $r_P = -0.23$; still negative in both cases, but closer to zero. The fact that these values are not positive invalidate the assumption we made to use this family of metrics to measure domainness. The results show how the density of the characteristic vocabulary is neither a sufficient nor a necessary condition to obtain in-domain corpora. It can be a good estimator for the representativity of the corpus, but if the cohesion is low, the domainness will be low too.

The additional normalization of this measure included in the augmented frequency $\hat{c}_{\text{terms}}$ rules out the metric as a global measure. The Pearson correlation for $\hat{c}_{\text{terms}}$ when all the data are used together is $r_P = -0.08$: these variables do not correlate. Since the term frequencies are now normalized to the most frequent term, their importance is lower, and therefore, both WT and IR behave similarly, with slightly higher values for IR. The reason is the same as before, exhibiting an anticorrelation with precision scores. However, when looking into the two systems, the correlation increases more for WT: $r_P = 0.63$ for WT and $r_P = 0.36$ for

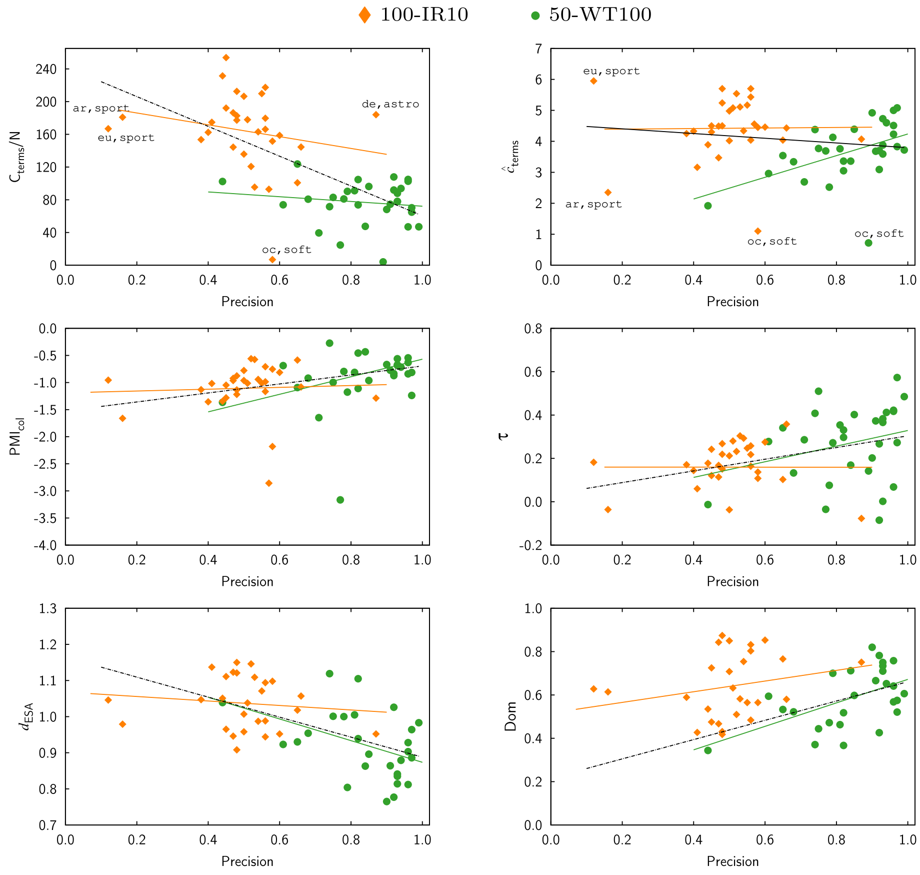


Fig. 6 Relation between six domainness measures and the precision given by human judgments (see text for correlations). Points correspond to the score for the 10 languages in the three manually evaluated domains, selected examples are highlighted

IR. So, within a system, we have a positive correlation of $\hat{c}_{\text{terms}}$ vs *precision*, indicating that $\hat{c}_{\text{terms}}$ is a good barometer of the quality of a WT-extracted in-domain corpus.

Family 2 Metrics related to mutual information or co-occurrence show a clear positive trend with respect to precision. Even with negative PMI values, human judgments show how the best collections have higher PMIs. The score that correlates best with precision is PMI$_{\text{col}}$ with $r_P = 0.57$. The metric with the standard probability calculation PMI$_{\text{art}}$ is close with $r_P = 0.55$. The variable-size sliding window of an article does not affect the results. The normalized versions lie slightly below because the normalization smooth the differences among points (NPMI$_{\text{art}}$ has $r_P = 0.41$; NPMI$_{\text{col}}$ has $r_P = 0.55$). In our setting, the median of (N)PMI is a better estimator than the mean.

When comparing the subset of points belonging to WT and IR, the correlation is lower than the global one in both cases, but specially for IR, where we observe no correlation between the metric and the observations (PMI$_{\text{art}}^{\text{WT}}$ has $r_P = 0.44$; PMI$_{\text{art}}^{\text{IR}}$ has $r_P = 0.08$). The different nature of WT and IR allows to say that a high density of in-domain terms in an article does not imply that it belongs to the domain, as concluded from the fact that C_{terms}/N

and $\hat{c}_{\text{terms}}$ for the IR system are above their WT equivalents. However, a higher number of co-occurrences of the domain vocabulary does ($\text{PMI}^{\text{WT}} > \text{PMI}^{\text{IR}}$).

Family 3 Metrics ρ and τ measure the rank correlation between the terms of an extracted in-domain collection and a collection of Wikipedia root articles in the same domain. The correlation with soft precision is $r_P = 0.31$ for ρ, and $r_P = 0.34$ for τ. As the plot for τ in Figure 6 shows, the dispersion of the WT points is larger, but their subset has a higher correlation than the IR one ($r_P = 0.25$ vs $r_P = 0.02$). For the IR subset, the metric is a very bad measure of the quality of the extraction, but contrary to the augmented term frequency metric $\hat{c}_{\text{terms}}$, it performs better in the global setting than within the subsets.

Family 4 ESA distances result in a good estimator for the cohesion of the corpus. With a global correlation of $r_P = -0.60$ and subset correlations of $r_P = -0.41$ (WT) and $r_P = -0.13$ (IR), d_{ESA} is the best individual metric to estimate the domainness of a collection in general, but $\hat{c}_{\text{terms}}$ is the best metric when we focus on WT extractions. $\hat{c}_{\text{terms}}$ is not bounded. Its range is $[0, \infty)$, where high densities imply a good quality. However, due to the lack of top boundary, it is useful to compare collections, but no clear interpretation exists in terms of an absolute number. In terms of ease of use, both d_{ESA} and $\hat{c}_{\text{terms}}$ rely on the Wikipedia. $\hat{c}_{\text{terms}}$ comes for free with a WT extraction because we estimate the characteristic vocabulary in our models. d_{ESA} performs better globally, but the cost is the need to define a reference collection, which can be different across languages. PMI_{col} alleviates this problem being also language independent, but its quality as a metric is slightly lower.

Finally, we estimate the domainness as the combination of the most promising metrics for representativity and cohesion:

$$\text{domainness} \equiv \text{Dom} = \left(\widehat{\text{PMI}_{\text{col}}} + \widehat{d_{\text{ESA}}}\right)/2, \tag{11}$$

where hats in $\widehat{\text{PMI}_{\text{col}}}$ and $\widehat{d_{\text{ESA}}}$ represent a normalization of the points in range [0,1]. As expected, we obtain the largest global correlation with the combination because representativity and cohesion are two perpendicular features. Dom reaches a correlation of $r_P = 0.71$ when all 60 datapoints are used. At system level, with two sets of 30 datapoints, Dom^{WT} has $r_P = 0.55$ and Dom^{IR} $r_P = 0.27$ showing that the more homogeneous a collection of points, the less important is the combination of aspects. This correlation is slightly worse than the one given by the simple augmented term frequency metric $\hat{c}_{\text{terms}}$, as seen before.

7 Summary and conclusions

Several multilingual applications benefit from in-domain corpora, but gathering them usually requires a considerable amount of work. We designed WikiTailor, a system to extract such corpora from the Wikipedia, a multilingual encyclopedia where the domain of an article is encoded in its category tags. WikiTailor explores Wikipedia's category graph and performs a breadth-first search departing from the category associated with the desired domain. From this point, it extracts all the articles belonging to its children categories down to an automatically estimated optimal depth. We compared the performance of WikiTailor with a standard IR system based on querying the Wikipedia with a set of keywords that describe the domain. The two methods are very different in nature and generate complementary collections with small intersections. Experiments on 10 languages and 743 domains showed the preference by automatic and manual evaluations for the WT models.

A crowdsourced manual evaluation was carried out on three domains—*Astronomy*, *Software*, and *Sport*—on one WT and one IR model. Turkers were asked to indicate if an article belonged to the domain or not, for a total of 200 articles per language and system. Precision was used to evaluate the quality of each collection. With average precisions of P^{WT}=0.84±0.13 and P^{IR}=0.50±0.14, WikiTailor resulted statistically better.

The lack of metrics to measure the domainness of a corpus made an automatic evaluation difficult. Therefore, we defined domainness as a combination of the representativity and coherence of the texts in a corpus and we introduced several metrics to account for it. Representativity is measured on the basis of the characteristic vocabulary of its intended domain (density, co-occurrence, or correlations between term distributions) and coherence on the basis of the distance between the articles of the collection. Via the correlation with human judgments, we show how the density of the characteristic vocabulary of the domain is neither a sufficient nor necessary condition for in-domain corpora. IR systems, with a higher density of in-domain terms by construction, are worse for all languages and domains in our manual evaluation. On the other hand, distances between the documents of a collection, as measured by explicit semantic analysis representations, outperform term-based measures and show a moderate correlation with observations.

Mathematically, we introduce the Dom metric: a normalized linear combination between the best representativity metric ($\widehat{\mathrm{PMI}_{\mathrm{col}}}$) and the distance-based one for coherence ($\widehat{d}_{\mathrm{ESA}}$). This combination shows a strong correlation with human evaluations, 0.71. In summary, d_{ESA} is the best individual metric to estimate the quality of a collection in general, when comparing heterogeneous collections as different in nature as the ones we explored. However, it is only measuring the coherence between the documents and the performance is improved when combined with a measure of the importance of in-domain term co-occurrences. Within a system, conclusions change. WT systems extract the articles without any request on the number of in-domain terms that the documents have, and within these collections the occurrences and co-occurrences of terms are relevant. For homogeneous collections, (WT *or* IR) $\hat{c}_{\mathrm{terms}}$ is the best metric. For heterogeneous collections, (WT *and* IR) d_{ESA} and Dom are the best options, meaning that coherence is more important when discrepancies in the number of in-domain vocabulary are not huge.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10115-022-01767-5.

Acknowledgements This work was partially funded by the TACARDI project (TIN2012-38523-C02) of the Spanish Ministerio de Economía y Competitividad (MEC). Lluís Màrquez's work on this paper was done while he was affiliated with the Qatar Computing Research Institute (QCRI).

Funding Open Access funding enabled and organized by Projekt DEAL.

A Wikipedia-specific concepts

Category	Tag present in a set of articles grouped together by covering similar topics.
Dump	Snapshot of an edition in the form of wikitext source and metadata embedded in XML.
Edition	Each one of the Wikipedias for a specific language.
Inter-language link/langlink	A link in a Wikipedia article toward an article on the same topic in a different language.
Main namespace	The namespace in the Wikipedia containing the actual contents: the articles. Other namespaces are user, help, or category.
WCG	Wikipedia category graph. Directed acyclic graph formed by the category tags.

B WikiTailor parameters

List of parameters for each of the methods (WT and IR) with the default value implemented in the `WikiTailor` toolkit. All parameters can be changed via configuration file or command line options. A more extense description can be found in the technical manual.[16]

Parameter	Description	Default	Method
Vocabulary definition (domain keywords)			
minNumArticles	Minimum number of articles in the root category to be considered for term extraction. If the root category has less articles, its children are also considered	10	WT, IR
topPercentage	Percentage of terms in the root vocabulary (that extracted from the root category articles) to be considered as in-domain terms	10	WT, IR
topKeywords	Number of vocabulary terms within the topPercentage to be considered. Since the percentage depends on a category itself and can grow a lot, this allows to use a fix number	100	WT, IR

[16] http://cristinae.github.io/WikiTailor/dwnld/wikiTailorTechnicalManual_v1.0.pdf.

Parameter	Description	Default	Method
Graph article selection			
minDepth	Minimum number of levels of subcategories from the desired root category to be considered in the graph exploration	0	WT
maxDepth	Maximum number of levels subcategories from the desired root category to be considered in the graph exploration	50	WT
percentage	Percentage of positive articles (articles with a domain keyword in the title) accepted for considering a level as in-domain	0.5	WT
depth	WT determines the optimal domain depth, but can be also fixed by command line	–	WT
IR article selection			
minPercentage	minimum Lucene relevance score for an article to be retrived as a percentage of the maximum score	10	IR

C Crowdsourcing settings

Setting up the Figure Eight crowdsourcing annotation involves four steps: (*i*) the selection of Turkers, (*ii*) their instruction, (*iii*) setting the task itself and (*iv*) a quality control of the annotation.

The selection of the Turkers was made by their language knowledge. We opted for three different criteria based upon language capabilities or region to determine the population that annotating each language. No language or geographical limitation was set for English, composing our most flexible configuration. For Arabic, French, and German, we selected the corresponding language on the platform interface. Such a setting was not available for the rest of languages;[17] hence, we opted for a geographical configuration. Table 7 summarizes the geographical configurations, set according to four criteria: countries where the language is official (e.g., Spain for Spanish), countries with official languages from the same family (e.g., France for Catalan), neighboring countries (e.g., Bulgaria for Greek), and countries with a high rate of immigration of native speakers (e.g., Germany for Greek).

We set the job as a binary classification task where Turkers had to assess if a Wikipedia article matches the domain displayed in the interface or not.

Instruction:

> **Task**
> - Identify the category a given Wikipedia article belongs to. It either belongs to domain *d* or to *other*, where *d* can be *Astronomy*, *Software*, or *Sport*.

[17] Spanish is an exception. In that case, we opted for shaping the demographics geographically, as speakers of Romance languages can often read contents in another Romance language.

Table 7 Geographical settings for the Figure Eight workers selection (when a language-based filtering was not available)

Language	Regions
Spanish	Spain, Portugal, Latin America, France
Romanian	Romania
Catalan	Andorra, Spain, Portugal, Latin America, France
Basque	France and Spain
Greek	Albania, Bulgaria, Cyprus, Greece, Macedonia, Turkey, Germany,USA
Occitan	France

The Turkers had to scroll an actual Wikipedia article, which we framed into the interface, to judge.

After a pilot experiment, we wrote additional specific guidelines for each of the three domains aiming at clarifying how some ambiguous cases should be handled by the annotators:

Astronomy
- The biography of an astronomer should be considered within the Astronomy domain.
- Articles about Physics should not always be considered as Astronomy even if atoms, particles or orbits are involved.

Software
- Concepts which are in essence software (e.g., video games, matchboxes) belong to the Software domain.

Sport
- The biography of a sportsman should be considered within the Sport domain.
- An article of a location with a section on Sport does not belong to the domain Sport.

We paid 0.06 USD per HIT. Each HIT consisted of 10 binary annotations and had a minimum working time of 120 seconds. We manually annotated 10% of the instances for quality control and requested an annotation accuracy of 80% to verify the annotation quality. Each item was judged three times.

References

1. Adafre S, de Rijke M (2006) Finding Similar Sentences across Multiple Languages in Wikipedia, In: Proceedings of the 11th conference of the European chapter of the association for computational linguistics (EACL), pp 62–69
2. Aker A, Kanoulas E, Gaizauskas R (2012) A light way to collect comparable corpora from the Web, In: Calzolari N, Choukri K, Declerck T, Dogan M, Maegaard B, Mariani J, Odijk J and Piperidis S (eds) Proceedings of the eighth international conference on language resources and evaluation (LREC), European Language Resources Association (ELRA), Istanbul, Turkey, pp 15–20
3. Artetxe M, Schwenk H (2019) Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Trans Assoc Comput Linguist (TACL) 7:597–610
4. Aspert N, Miz V, Ricaud B, Vandergheynst P (2019) A Graph-Structured Dataset for Wikipedia Research, In: Companion Proceedings of The 2019 World Wide Web conference (WWW), Association for Computing Machinery (ACM), New York, NY, USA, pp 1188–1193

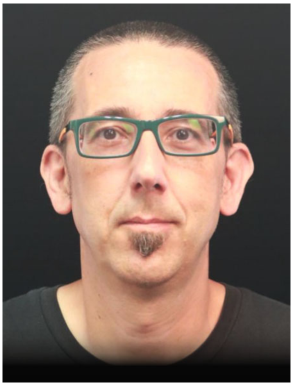
5. Barrón-Cedeño A, España-Bonet C, Boldoba J, Màrquez L (2015) A Factory of Comparable Corpora from Wikipedia, In: Proceedings of the 8th Workshop on Building and Using Comparable Corpora (BUCC), Beijing, China, pp 3–13. http://www.aclweb.org/anthology/W15-3402
6. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist (TACL) 5:135–146
7. Bouma G (2009) Normalized (Pointwise) mutual information in collocation extraction, In: Proceedings of the Biennial GSCL conference, Tübingen, Germany, pp 31–40
8. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J D, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners, In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, and Lin H, (eds) Advances in Neural Information Processing Systems, Vol 33, Curran Associates, Inc., pp 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
9. Cheon J, Ko Y (2021) Parallel sentence extraction to improve cross-language information retrieval from Wikipedia. J Inf Sci 47(2):281–293. https://doi.org/10.1177/0165551521992754
10. Chi Z, Dong L, Ma S, Huang S, Singhal S, Mao X-L, Huang H, Song X, Wei F (2021) mT6: multilingual pretrained text-to-text transformer with translation pairs, In: Proceedings of the 2021 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 1671–1683. https://aclanthology.org/2021.emnlp-main.125
11. Chu C, Nakazawa T, Kurohashi S (2014) Iterative Bilingual Lexicon Extraction from Comparable Corpora with Topical and Contextual Knowledge. In: Gelbukh A (ed) Computational Linguistics and Intelligent Text Processing, vol. 8404. Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp 296–309
12. Croux C, Dehon C (2010) Influence unctions of the Spearman and Kendall correlation measures. Stat Methods Appl 19(4):497–515
13. Cui G, Lu Q, Li W, Chen Y (2008) Corpus Exploitation from Wikipedia for Ontology Construction, In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odjik J, Piperidis S, and Tapias D (eds) Proceedings of the sixth international language resources and evaluation (LREC), European Language Resources Association (ELRA), Marrakech, Morocco, pp 2126–2128
14. Darwish K, Magdy W (2014) Arabic information retrieval. Now Publishers Inc, Hanover MA, Foundations and Trends
15. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding, In: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers). (NAACL-HLT), Association for Computational Linguistics (ACL), Minneapolis, Minnesota, pp 4171–4186. https://www.aclweb.org/anthology/N19-1423
16. Erdmann M, Nakayama K, Hara T, Nishio S (2008) An Approach for Extracting Bilingual Terminology from Wikipedia, In: Proceedings of the 13th international conference on database systems for advanced applications (DASFAA), Springer-Verlag, Berlin, Heidelberg, pp 380–392. http://dl.acm.org/citation.cfm?id=1802514.1802552
17. España-Bonet C, Barrón-Cedeño A, Màrquez L (2020) WTC1.1 (WikiTailor Corpus v.1.1)
18. Esplá-Gomis M, Forcada ML (2009) Bitextor, a Free/Open-Source Software to Harvest Translation Memories from Multilingual Websites, In: Gerber L (ed) Beyond translation memories: new tools for translators workshop, Ottawa, Canada. https://aclanthology.org/2009.mtsummit-btm.6
19. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378–382
20. Gabrilovich E, Markovitch S (2007) Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis, In: Proceedings of the 20th international joint conference on artificial intelligence (IJCAI), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 1606–1611
21. Gamallo Otero P, González López I (2010) Wikipedia as Multilingual Source of Comparable Corpora, In: Proceedings of the 3rd Workshop on Building and Using Comparable Corpora (BUCC), pp 21–25
22. Gamallo Otero P, González López I (2011) Measuring Comparability of Multilingual Corpora Extracted from Wikipedia. In: Rosso P, Barrón-Cedeño A, Vila M, Civera J, Barreiro A, Alegria I (eds) Workshop on Iberian Cross-Language NLP tasks. Huelva, Spain, pp 8–13
23. Gottron T, Anderka M, Stein B (2011) Insights into Explicit Semantic Analysis, In: Proceedings of the 20th ACM international conference on information and knowledge management (CIKM), Association for Computing Machinery (ACM), Glasgow, Scotland, pp 1961–1964
24. Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T (2018) Learning Word Vectors for 157 Languages, In: Proceedings of the eleventh international conference on language resources and evaluation (LREC), European Language Resources Association (ELRA), Miyazaki, Japan, pp 3483–3487

25. Harsha Ramesh S, Prasad Sankaranarayanan K (2018) Neural Machine Translation for Low Resource Languages using Bilingual Lexicon Induced from Comparable Corpora, In: Annual conference of the North American chapter of the association for computational linguistics (NAACL) Student Research Workshop, pp 112–119. https://www.aclweb.org/anthology/N18-4016
26. Hassan S, Mihalcea R (2009) Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge, In: Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 3 (EMNLP), Association for Computational Linguistics (ACL), Singapore, pp 1192–1201. http://dl.acm.org/citation.cfm?id=1699648.1699665
27. Hecht B, Gergle D (2010) The Tower of Babel Meets Web 2.0: user-generated content and its applications in a multilingual context, In: Proceedings of the SIGCHI conference on human factors in computing systems, Association for Computing Machinery (ACM), Atlanta, GA, pp 291–300
28. Jakubina L, Langlais P (2016) A comparison of methods for identifying the translation of words in a comparable corpus: recipes and limits. Computación y Sistemas 20:449–458
29. Kilgarriff A (2001) Comparing corpora. Int J Corp Linguist 6(1):1–37
30. Lample G, Conneau A, Ranzato M, Denoyer L, Jégou H (2018) Word Translation without Parallel Data, In: Proceedings of the 6th international conference on learning representations, ICLR, Vancouver, Canada. https://openreview.net/forum?id=H196sainb
31. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174
32. Le Q, Mikolov T (2014) Distributed Representations of Sentences and Documents, In: Proceedings of the 31st international conference on machine learning—Volume 32 (ICML), JMLR, Beijing, China, pp 1188–1196
33. Ma S, Dong L, Huang S, Zhang D, Muzio A, Singhal S, Awadalla HH, Song X, Wei F (2021) DeltaLM: encoder-decoder Pre-training for Language Generation and Translation by Augmenting Pretrained Multilingual Encoders, CoRR arXiv:2106.13736
34. Munteanu D, Marcu D (2005) Improving machine translation performance by exploiting non-parallel corpora. Comput Linguist 31(4):477–504
35. Newman D, Lau JH, Grieser K, Baldwin T (2010) Automatic Evaluation of Topic Coherence. In: Kaplan R, Burstein J, Harper M, Penn G (eds) Human Language Technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics (NAACL-HLT). Association for Computational Linguistics (ACL), Los Angeles, CA, pp 100–108
36. Pinnis M, Ion R, Ştefănescu D, Su F, Skadiņa I, Vasiļjevs A, Babych B (2012) , ACCURAT Toolkit for Multi-level Alignment and Information Extraction from Comparable Corpora, In: Proceedings of the annual meeting of the association for computational linguistics (ACL) 2012 System Demonstrations, Association for Computational Linguistics (ACL), Jeju Island, Korea, pp 91–96. http://dl.acm.org/citation.cfm?id=2390470.2390486
37. Plamada M, Volk M (2012) Towards a Wikipedia-Extracted Alpine Corpus, In: Proceedings of the 5th workshop on building and using comparable corpora: language resources for machine translation in less-resourced languages and domains (BUCC), Istanbul, Turkey, pp 81–87
38. Plamada M, Volk M (2013) Mining for Domain-Specific Parallel Text from Wikipedia, In: Proceedings of the sixth workshop on building and using comparable corpora (BUCC), Association for Computational Linguistics (ACL), Sofia, Bulgaria, pp 112–120. http://www.aclweb.org/anthology/W13-2514
39. Porter M (1980) An algorithm for suffix stripping. Program 14:130–137
40. Potthast M, Stein B, Anderka M (2008) A Wikipedia-Based Multilingual Retrieval Model, In: Advances in information retrieval, 30th European conference on IR research (ECIR)
41. Pouliquen B, Steinberger R, Ignat C (2003) Automatic Identification of Document Translations in Large Multilingual Document Collections, In: Proceedings of the international conference recent advances in natural language processing (RANLP), Borovets, Bulgaria, pp 401–408
42. Prochasson E, Fung P (2011) Rare Word Translation Extraction from Aligned Comparable Documents, In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies—Volume 1 (HLT), Association for Computational Linguistics (ACL), Portland, OR, pp 1327–1335
43. Resnik P, Smith N (2003) The web as a parallel corpus. Comput Linguist 29(3):349–380
44. Ruiter D, España-Bonet C, van Genabith J (2019) Self-Supervised Neural Machine Translation, In: Proceedings of the 57th annual meeting of the association for computational linguistics, Volume 2: Short Papers. (ACL), Association for Computational Linguistics (ACL), Florence, Italy, pp 1828–1834. https://www.aclweb.org/anthology/P19-1178
45. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manag 24(5):513–523

46. Schwenk H, Chaudhary V, Sun S, Gong H, Guzmán F (2021) WikiMatrix: mining 135M parallel sentences in 1620 language pairs from Wikipedia, In: Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume (EACL), Association for Computational Linguistics (ACL), Online, pp 1351–1361. https://aclanthology.org/2021.eacl-main.115
47. Skadiņa I, Aker A, Mastropavlos N, Su F, Tufi D, Verlic M, Vasiļjevs A, Babych B, Clough P, Gaizauskas R, Glaros N, Paramita ML, Pinnis M (2012) Collecting and Using Comparable Corpora for Statistical Machine Translation, In: Calzolari N, Choukri K, Declerck T, Dogan M, Maegaard B, Mariani J, Odijk J, and Piperidis S (eds) Proceedings of the eighth international conference on language resources and evaluation (LREC), European Language Resources Association (ELRA), Istanbul, Turkey
48. Smith JR, Quirk C, Toutanova K (2010) Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In: Kaplan R, Burstein J, Harper M, Penn G (eds) Human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics (NAACL-HLT). Association for Computational Linguistics (ACL), Los Angeles, CA, pp 403–411
49. Ştefănescu D, Ion R, Hunsicker S (2012) Hybrid Parallel Sentence Mining from Comparable Corpora, In: Proceedings of the 16th annual conference of the European association for machine translation (EAMT), European Association for Machine Translation, Trento, Italy, pp 137–144. https://aclanthology.org/2012.eamt-1.37
50. Su F, Babych B (2012) Measuring Comparability of Documents in Non-parallel Corpora for Efficient Extraction of (Semi-)Parallel Translation Equivalents, In: Proceedings of the joint workshop on exploiting synergies between information retrieval and machine translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), EACL 2012, Association for Computational Linguistics (ACL), Avignon, France, pp 10–19
51. Talvensaari T, Pirkola A, Järvelin K, Juhola M, Laurikkala J (2008) Focused Web crawling in the acquisition of comparable corpora. Inf Retr 11(5):427–445
52. Tao T, Zhai C (2005) Mining Comparable Bilingual Text Corpora for Cross-Language Information Integration, In: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, Association for Computing Machinery (ACM), New York, NY, pp 691–696
53. Varga A (2017) Domain Adaptation for Multilingual Neural Machine Translation, Master's thesis, Saarland University
54. Vu T, Aw ATT, Zhang M (2009) Feature-based Method for Document Alignment in Comparable News Corpora, In: Lascarides A, Gardent C, and Nivre J (eds) Proceedings of the 12th conference of the European chapter of the association for computational linguistics (EACL), Association for Computational Linguistics (ACL), Athens, Greece, pp 843–851
55. Yasuda KT, Sumita E (2008) Method for Building Sentence-Aligned Corpus from Wikipedia, In: Proceedings of the AAAI workshop on wikipedia and artificial intelligence: an evolving synergy, Menlo Park, CA, pp 64–66
56. Yu KT, Tsujii J (2009) Bilingual Dictionary Extraction from Wikipedia, In: Gerber L (ed) Proceedings of MT Summit XII, Ottawa, Canada
57. Zesch TT, Gurevych I (2007) Analysis of the Wikipedia Category Graph for NLP Applications, In: Proceedings of the second workshop on textgraphs: graph-based algorithms for natural language processing, Association for Computational Linguistics (ACL), Rochester, NY, USA, pp 1–8. https://www.aclweb.org/anthology/W07-0201
58. Zesch T, Müller CT, Gurevych I (2008) Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary, In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odjik J, Piperidis S, and Tapias D (eds) Proceedings of the sixth international language resources and evaluation (LREC), European Language Resources Association (ELRA), Marrakech, Morocco, pp 1646–1652. http://www.lrec-conf.org/proceedings/lrec2008/pdf/420_paper.pdf

Cristina España-Bonet is a senior researcher team lead at the German Research Center for Artificial Intelligence (DFKI GmbH) in Saarbrücken, Germany. She received the B.E. in physics and the M.Sc. in astrophysics and cosmology from the Universitat de Barcelona (Catalonia) in 2002 and 2004, respectively. In 2008, she obtained the M.Sc. in artificial intelligence from the Universitat Politècnica de Catalunya (Catalonia) and the Ph.D. in physics from the Universitat de Barcelona. Since then, she has been working on natural language processing and deep learning first at Universitat Politècnica de Catalunya and currently at DFKI and the Universität des Saarlandes. Cristina has participated in 16 competitive funded projects, has 60+ peer reviewed publications and organized 6 scientific workshops and 3 shared tasks. She has been teaching in the Physics, Computer Science and Computational Linguistics faculties and advised 16 Master and 4 PhD students. She is especially interested in interlingual, multilingual and multimodal approaches and in making available tools and methods for low-resourced settings and languages.

Alberto Barrón-Cedeño is currently Associate Professor (Reader) of Natural Language Processing in the Department of Interpreting and Translation at Università di Bologna (Italy). Previous to that, he spent 4.5 years as Research Scientist at Qatar Computing Research Institute (Qatar) and 2 years in TALP Research Centre at Technical University of Catalonia-BarcelonaTech (Spain). With main research interests in natural language processing, information retrieval and machine learning, he has authored over 70 technical papers. Dr Barrón-Cedeño has served as the general chairman of CLEF 2022 and as the organiser of diverse shared tasks and events, such as CheckThat! at CLEF, PAN at CLEF, and SemEval, among others. He received the B.E. degree in computing and the M.Sc. degree in computing science from the National University of Mexico, Mexico, in 2004 and 2007, respectively, and the M.Sc. and Ph.D. degrees in computing science from the Technical University of Valencia, Spain, in 2008 and 2012.



Lluís Màrquez is a Principal Applied Scientist at the AWS AI Labs from Amazon in Barcelona. He joined Amazon in January 2018. Before that, Lluís spent 20 years in Academia (as an Associate Professor at the Polytechnical University of Catalonia, UPC) and 5 years at the Qatar Computing Research Institute in Doha (2013-2018). His research during that time focused on natural language understanding by using statistical machine learning models. He worked extensively on syntactic and semantic parsing, with applications to machine translation, Question and Answering, and information verification. He has 160+ publications in Natural Language Processing (NLP) and Machine Learning journals and conferences. He took several distinguished roles in the NLP community including: General/Program Chair of major conferences in the area (ACL, EMNLP, EACL, CoNLL, *SEM, and EAMT), President of the Association for Computational Linguistics (ACL) special interest group on Natural Language Learning (SIGNLL; 2009-2011), President of the European Chapter of the ACL (EACL; 2013-2016), and member of the ACL Executive Committee (2015-2016). He has participated in numerous Spanish and EU research and technology-transfer projects throughout his career.