Departamento de Inteligencia Artificial
Escuela Técnica Superior de Ingenieros Informáticos

PhD Thesis

Semantically-enabled browsing of large multilingual document collections

Author: Carlos Badenes-Olmedo
Supervisors: Prof. Dr. Oscar Corcho

xxx, 2020

Tribunal nombrado por el Sr. Rector Magfco. de la Universidad Politécnica de Madrid, el día XX de xxx de 2020.

Presidente:	Dr. Xxx Xxx
Vocal:	Dra. Xxx Xxx
Vocal:	Dr. Xxx Xxx
Vocal:	Dra. Xxx Xxx
Secretario:	Dr. Xxx Xxx
Suplente:	Dra. Xxx Xxx
Suplente:	Dr. Xxx Xxx

Realizado el acto de defensa y lectura de la Tesis el día X de xxx de 2020 en la Escuela Técnica Superior de Ingenieros Informáticos

Calificación: ____________________

EL PRESIDENTE VOCAL 1 VOCAL 2

VOCAL 3 EL SECRETARIO

A mis padres.
A Beatriz.
A Martín y Alonso.

Agradecimientos

xxxxx

Abstract

xxxxx

Resumen

xxxxxx

Contents

List of Figures

List of Tables

Acronyms

API: Application Programming Interface

CQ: Competency Question

GUI: Graphical User Interface

IDE: Integrated Development Environment

LD: Linked Data

LOD: Linked Open Data

UML: Unified Modeling Language

URI: Uniform Resource Identifier

URL: Uniform Resource Locator

WUI: Web User Interface

Chapter 1

Introduction

Explicar bien: - thematic associations - representational models (vector space models)

1.1 Contributions

..

1.2 Thesis Structure

..

1.3 Publications

..

Chapter 2

Related Work

2.1 Text Processing

2.2 Document Embeddings

2.3 Document Similarity

..

2.4 Summary

..

Chapter 3

Research Objectives

The work presented in this thesis aims to facilitate the exploration of large collections of multilingual documents through thematic associations inferred from their content. Each of the challenges arising from this objective defines a working dimension and guides the research carried out in this thesis.

The first dimension focuses on **scalability** creating the texprocessing flows that are usually required to create or apply learning models. The workload required to process a corpus varies according to the number of documents, the length of texts and the model to be created. If the design of the workflow is scalable, there is no need to modify the processing logic when working with larger collections of documents, since adding more resources is enough to perform it. These resources can be machines (i.e horizontal scaling) or processing units (e.g CPU, RAM) in an existing machine (i.e vertical scaling).

The second dimension covers the **representativeness** of the text representations when projected into spaces where they are manipulated. The idea behind these spaces is to represent documents as points (or vectors in a vector space) that are close together when the texts are semantically similar, and far apart when they are semantically distant. The ability of these spaces to create meaningful representations is studied in this work.

In the third dimension, the **discriminatory** capacity of the structures used to distribute texts from their representations is studied. The content of documents is analyzed to relate them and allow large collections being browsed through their relations.

The *representativeness* covered in the previous dimension enables the interpretation of the relations obtained.

And finally, the fourth dimension handles the **multilingualism** of collections with documents in several languages. On a multilingual space, documents are described and related across languages.

This chapter introduces our main hypothesis (3.1), its associated research challenges (3.2) and presents the research methodology (3.3).

3.1 Research Hypotheses

We define our main hypothesis as follows:

Hypothesis 1 *Large multilingual document collections can be automatically analyzed to discover the thematic relations that facilitate a semantically-enabled text browsing.*

Our hypothesis can be divided into four different sub-parts, which are related to the aforementioned scalability, representativeness, discrimination, and multilinguality dimensions respectively. First, by *distributing both natural language processing tasks and representational models we can efficiently process big collections of documents(**H1.1**).*

Second, we can *semantically relate documents by comparing their most relevant topics (**H1.2**).* Furthermore, for this purpose we hypothesize that the use of *topic hierarchies (**H1.2.1**) and similarity metrics based on relevance levels(**H1.2.2**) can quantify the semantic distance between texts.* Third, by *dividing the representational space in regions based on topics and relevance levels we can search for related documents without having to calculate all pairwise comparisons and without losing the ability to filter through their topics (**H1.3**).*

And finally, *by abstracting the topic representations into concept-based descriptions across languages we can relate documents in various languages without having to translate them (**H1.5**).*

A summary of the hypotheses and how they deal with our research objectives can be found in table 3.1.

Hypothesis	Research Dimension
H1: Large multilingual document collections can be automatically analyzed to be semantically-browsed through thematic relations	D1: Scalability, D2: Representativeness, D3: Discrimination, D4: Multilingualism
H1.1: it is possible to efficiently annotate documents on a large scale by distributing natural language processing tasks and representation models	D1: Scalability
H1.2: it is possible to semantically relate texts from their most relevant topics	D2: Representativeness
H1.3: it is possible to find documents with similar topic distributions without calculate all pairwise comparisons and without losing the ability to explore them through their topics	D2: Representativeness, D3: Discrimination
H1.4: it is possible to relate documents in different languages without having to translate them using language agnostic concepts from their main topics	D4: Multilingualism

Table 3.1: Hypotheses and research dimensions.

3.2 Research Challenges

Several research challenges emerge from these hypotheses. First, in order to facilitate reusing existing topic models by processing systems with different architectures and technological stacks, we have to work on *topic-model programming interfaces*. Second, in order to describe and thematically relate documents, we must address *explainable topic-based associations*. Third, by working with huge collections of documents described by topics, we need to handle *large-scale comparisons of topic distributions*. Finally, in order to explore multilingual document collections from shared topic-based representational spaces, we have to provide *automatic cross-lingual topic alignment*. Each of these research challenges are described below and covered throughout the thesis.

3.2.1 Topic-model Programming Interface

Although some initiatives exist to standardize the format of machine-learning models and to provide tools that facilitate their transformation among the most widespread proprietary formats, there are still some software restrictions that can limit their reuse. These models may require a software dependency that forces using a specific version of a programming language (e.g., python2 vs python3[1]) or an operating system (e.g., linux kernel vs on-cloud environments[2]) to load them or to launch the service that deploys them (e.g., ONNX[3]). This limits their ability to be reused from domains non-familiar with these technological stacks. *Integrating pre-trained topic models into general-purpose systems is not easy **(RCInterface1)**.*

Topic models, as many other machine learning models, can then be distributed in a proprietary or standard format with software dependencies or directly by providing the data. However, *there is no standard presentation of the topics and the operations that can be performed on them. **(RCInterface2)***. Sometimes topics are distributed as a list of words with the top ten or five most relevant, and occasionally these word distributions are not even accompanied by weights, making any analysis of their densities impossible. On other cases, it is not even possible to infer the topic distributions in new texts, even when the algorithm used to generate the model allows it.

[1] `https://www.python.org`
[2] `https://vespa.ai`
[3] `https://onnx.ai`

3.2.2 Explainable Topic-based Associations

In order to facilitate the exploration of document collections, vector space models are often used to semantically relate texts based on their word distributions. In large collections, these models have been adapted to reduce the dimensions of the vectors and make the operations more manageable. As a result, a new abstraction named topics emerged between words and texts that also provides a particular knowledge of the whole collection. Topics are described by word distributions over the entire vocabulary and documents by vectors containing topic distributions. Despite the extensive use of these representation models, *there is no common criteria for identifying the most representative topics in a document **(RCExplainable1)***.

In addition, since similarity metrics in this representation space are based on accumulating the difference in topic densities, *it is difficult to understand the distance between topic distributions **(RCExplainable2)***. And, unless a minimum distance threshold is defined or a n-top topics agreed, *there is no common criterion for determining whether two documents are related**(RCExplainable3)***.

3.2.3 Large-scale Comparisons of Topic Distributions

There are many scenarios where finding related documents in a large corpus is desirable (e.g. a researcher doing literature review, or an R&D manager analyzing project proposals). Experts can benefit from discovering those connections to achieve those goals, but brute-force pairwise comparisons are not computationally adequate when the size of the corpus is too large. Some algorithms in the literature divide the search space into regions containing potentially similar documents, which are later processed separately from the rest in order to reduce the number of pairs compared. However, *there are no mechanisms that efficiently partition the topic-based search space without compromising the ability for thematic exploration **(RCComparison1)***.

In addition, documents from the same region should be compared and *there are no similarity metrics that compare partial distributions of topics **(RCComparison2)***.

3.2.4 Automatic Cross-lingual Topic Alignment

With the ongoing growth in the number of digital articles in different languages, we need annotation methods that enable browsing multi-lingual corpora. Multilingual probabilistic topic models have recently emerged as a group of semi-supervised machine learning models that can be used to perform thematic explorations on collections of texts in multiple languages. However, *there are no approaches that abstract the representation of probabilistic topics in language-independent spaces without translating texts or aligning documents* ***(RCCrossLingual1)***. Existing approaches require parallel or comparable training data to create a language-independent space.

A summary of the challenges covered in this work and how they map to the hypotheses is presented in table3.2

3.3 Research Methodology

The research presented in this thesis is based on four dimensions or research areas. Each one is motivated by a research problem that arises when performing a task that must be solved. Once a dimension is tackled, the next one is reached, and so on. This iterative and incremental methodology allows us refining the research results by evaluating them with more experiments and addressing increasingly complex research problems.

Figure 3.1 shows the dimensions on which the research of this thesis has been built. The top of the pyramid is only reached once the lower dimensions are dealt with. They are presented as a chain of four steps. The first step describes the motivation to perform a given task coming from real-world problems that we had to deal with and is represented by a brown arrow. In the context of this task, the research problem arises and is framed by a pink arrow. For each of them a solution is proposed and evaluated according to a specific criterion. The proposed solution is represented by a green arrow and the evaluation with a blue arrow. Once a proposal has been validated, the next dimension of the pyramid is achievable and all the previous research problems are added to the new research problem as conditions to be taken into account

Technical objectives (i.e., develop a new resource) or research objectives (i.e., discover the solution to a problem) may appear to guide the solution proposal before moving on to the next dimension. They are presented below, organized by the research problem associated with each dimension.

Research Challenge	Hypotheses
RCInterface1: integrating pre-trained topic models into general-purpose systems is not easy	H1.1: documents can be efficiently annotated on a large scale by distributing natural language processing tasks and representation models
RCInterface2: there is no standard presentation of topics that facilitates their reuse	H1.1: documents can be efficiently annotated on a large scale by distributing natural language processing tasks and representation models
RCExplainable1: there is no common criteria for identifying the most representative topics in a document	H1.2: texts can be semantically related from their most relevant topics, H1.3: documents with similar topic distributions can be found without calculate all pairwise comparisons and without losing the ability to explore them through their topics
RCExplainable2: it is difficult to understand the distance between topic distributions	H1.2: texts can be semantically related from their most relevant topics
RCExplainable3: there is no common criterion for determining whether documents are related	H1.2: texts can be semantically related from their most relevant topics
RCComparison1: there are no mechanisms that efficiently partition the topic-based search space without compromising the ability for thematic exploration	H1.3: documents with similar topic distributions can be found without calculate all pairwise comparisons
RCComparison2: there are no similarity metrics that compare partial distributions of topics	H1.3: documents with similar topic distributions can be found without calculate all pairwise comparisons
RCCrossLingual1: there are no approaches to abstract probabilistic topics in language-independent spaces without translating texts or aligning documents	H1.4: documents in different languages can be related without having to translate them using language agnostic concepts from their main topics

Table 3.2: Open Research Challenges and Hypotheses.

3.3.1 Scalable Creation and Inference of Topics

This first dimension arose when we had to analyze a huge collection of documents describing research and innovation projects to discover which research areas are being addressed, measure their presence in the collection, and characterize them so that they can be assigned to new documents. Such a high volume of data makes difficult to process it manually, so we need to automatize the required processing to draw insight from it. Probabilistic topics allow describing research areas so we define a *distributed text-processing model for creating large probabilistic topic models (**RO1**)* and a *web service template to distribute them (**RO2**)*. In this way, the models themselves can be easily integrated into scalable processing designs. As a result, we create a *platform for large-scale text analysis (**TO1**)*, and produce a *model-as-a-service repository with pre-trained topic models(**T02**)*. Its efficiency is validated by processing a corpus of 100,000 documents collected from the CORDIS dataset, which contains descriptions of projects funded by the European Union under a framework programme since 1990 (Badenes-Olmedo et al., 2017b).

The main contributions in this dimension are described in section 4 as follows:

- a software architecture to process big volumes of textual documents in a distributed and decoupled manner;
- a definition of model-as-a-service from the probabilistic topic models;
- an implementation of the architecture, librAIry, following those design principles;

3.3.2 Explainable Topic-based Associations

In the second dimension we needed to browse scientific papers through their content-based relations. The problem of massively annotating documents with topic distributions comes up. We have to *create annotations based on topic models in a way that is computationally affordable and enables a semantic-aware exploration of the knowledge inside it (**RO3**)*. Once documents are annotated, a *metric that compares documents and facilitates their interpretation from topic annotations (**RO4**)* is required. As a result, we integrate *the annotation method into the topic model service (**TO3**)* and implement a text comparison metric based on partial representations of topics. These proposals are validated by classifying 500,000 scientific articles from domains such as Computer Science, Neuroscience and Biomedicine (Badenes-Olmedo et al., 2017c) (Badenes-Olmedo et al., 2017a) (Badenes-Olmedo et al., 2019a).

The main contributions in this dimension are described in section5as follows:

- a clustering algorithm based on probabilistic topic distributions;
- a hash function to transform topic distributions into topic hierarchies;
- a similarity metric based on topic sets;

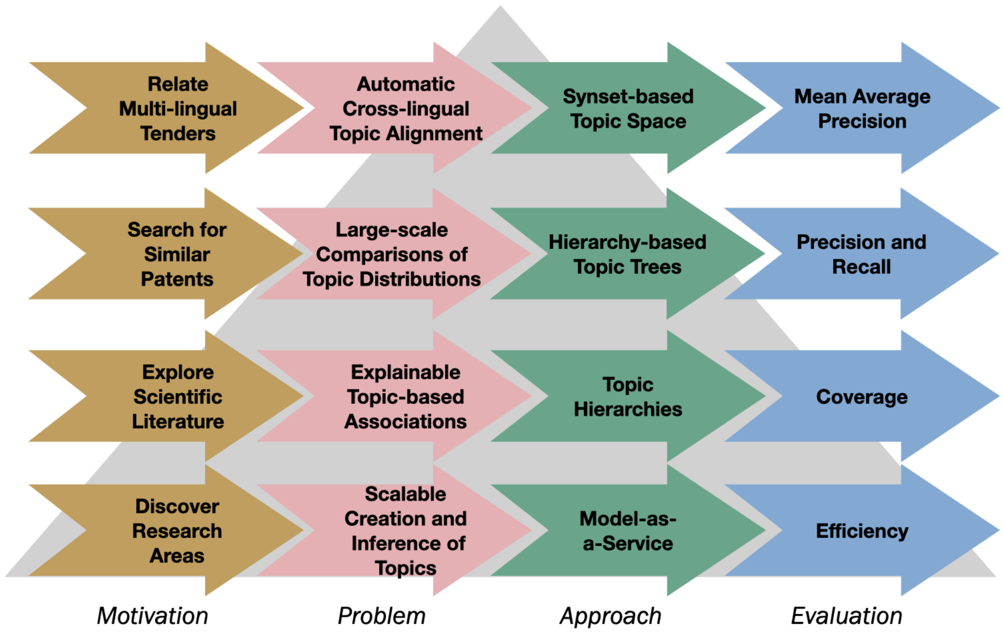

Figure 3.1: Research dimensions of the thesis. The first ones must be overcome before reaching higher dimensions.

3.3.3 Large-scale Comparisons of Topic Distributions

This dimension covers the search for similar documents based on their most relevant topics. Thanks to the above two dimensions, large collections of documents can be annotated with topic hierarchies and text distances can be measured from their annotations. Now, the aim is to find similar documents without losing the exploratory capacity offered by topics. Similarity comparisons are too costly to be performed in such huge collections of data and require more efficient approaches than having to calculate all pairwise similarities. We apply *techniques based on approximate nearest-neighbors to organize documents in regions with similar topic hierarchies (**RO5**)*. As a result, we develop *a system to automatically find similar documents (**TO4**)*. It is validated by comparing the manually related patents with those obtained automatically from their topic distributions in a collection of one million documents (Badenes-Olmedo et al., 2020)(Badenes-Olmedo et al., 2019a).

The main contributions in this dimension are described in section 6 as follows:

- a space-partitioning data structure to organize documents described by topic hierarchies
- a corpus browser that leverages these representations to automatically relate documents

3.3.4 Automatic Cross-lingual Topic Alignment

Finally, a new dimension on top of the previous ones emerges to relate texts coming from different languages. In particular, since document relations are based on their topics, this dimension is focused on aligning topics without supervision from models trained with texts in different languages. Since each language defines its own vocabulary, the topics are model-specific and cannot be directly compared. We *abstract the topic representations to create a unique space out of the particularities of the language (**RO6**)*.This approach is validated on the English, Spanish, French, Italian and Portuguese editions of JCR-Acquis corpora revealing promising results on classifying and sorting documents by similar content across languages (Badenes-Olmedo et al., 2019b)(Badenes-Olmedo et al., 2019a).

The main contributions in this dimension are described in section7, as follows:

- an algorithm to represent probabilistic topics using concept sets
- aligned topic models from the English, Spanish, French, Italian and Portuguese editions of the JRC-Acquis corpus

Table 3.3 summarizes the research objectives (ROs), technical objectives (TOs) and the research challenges (RCs) from Table 3.2 that they tackle.

Research Objective	Research Challenge
RO1: Define a distributed text-processing model for creating large probabilistic topic models	RCInterface1
RO2: Define a template to package probabilistic topic models as web services	RCInterface2
RO3: Define annotations based on topics that enables a semantic-aware exploration of the knowledge inside a corpus	RCExplainable1
R04: Define a metric based on topic annotations that compares documents and facilitates their interpretation	RCExplainable2, RCExplainable3
RO5: Define nearest-neighbors techniques to organize documents in regions with similar topic hierarchies	RCComparison1, RCComparison2
R06: Define a transformation of the topic-based annotations to create a unique representational space out of the particularities of each language	RCCrossLingual1
TO1: Create a platform for large scale text processing	RCInterface1, RCInterface2
T02: Create a repository of Topic-based web services	RCInterface2
T03: Integrate the annotation method based on topic hierarchies into the topic model service	RCExplainable2, RCComparison2
T04: Create a system capable of finding similar document automatically	RCExplainable2, RCExplainable3, RCComparison1, RCCrossLingual1

Table 3.3: Research and technical objectives and their related challenges.

Chapter 4

Scalable Creation and Inference of Topics

This chapter presents *librAIry*[4], our text management platform that combines natural language processing techniques with automatic learning algorithms to analyze large collections of documents. It serves as a technological framework where we can implement the advances of our research and measure its performance.

4.1 Document Workflow

Given the huge amount of textual data about any domain that is daily being produced or captured in any imaginable domain, it becomes crucial to provide mechanisms for programmatically processing this raw data so we can make sense out of it: discarding all the noisy, non-relevant information and keeping only the data that can bring value for the involved agents (general consumers, experts, companies, investors...).

While some specific tools already allow for advanced sense-making operations, others opt for composing a solution where different analysis techniques are integrated under a uniform data schema. However, this integration involves significant efforts on reconciling data sources, coordinating processing operations, and efficiently exploiting results from the execution of those techniques. There is the need for a more flexible paradigm where tools and algorithms for textual document analysis, from different programming languages and technologies, can operate independently and in a collaborative manner

[4] http://librairy.linkeddata.es

creating a common document oriented workflow through their actions. In the context of the scientific publications, the personalized recommendation of research papers based on their content is a key novel feature for performing a smart selection of relevant resources over very big collections of scientific content. From the set of values and different attributes extracted from the papers and by generating advanced knowledge models about the information they contain we can bridge across the different relevant pieces of information and allow users to navigate them in a more efficient and powerful way. This knowledge about a specific document is frequently acquired by different techniques focused on revealing certain aspects of it, that are later combined to achieve one particular task.

The architecture presented in this thesis aims to ease the way different software modules work together and lays the foundation for efficiently process big volumes of textual documents in a distributed, decoupled manner.

4.2 librAIry

librAIry is a framework where different text mining tools, available in various languages and technologies, can operate in a distributed, high-performance and isolated manner creating a common workflow through their actions. Instead to work towards a predefined sequence of actions, synchronization across modules is achieved through the aggregation of the operations executed by them in response to an emergent chain of events. This raises both technical and functional challenges to coordinate multiple executions. From the technical point of view, isolated environments and communication mechanisms are provided so initially dissimilar tools can be executed with maximum guarantees. From the functional point of view, all executions are coordinated to reach a final result as aggregation of partial results derived from each execution.

4.2.1 Functional Features

The architecture is articulated around three main concepts: (1) the **resource** such as *document*, a *part-of-a-document*, or a *domain*. (2) the **actions** performed over them: *create*, *update* or *delete* a resource. And (3) the new **state** that is reached by the resource after an action is performed, such as *created*, *updated* or *deleted*. An **event** is a message containing details about those three aspects, published on a shared

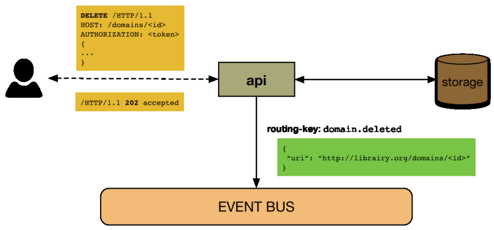

Figure 4.1: Domain deleted flow.

event-bus available for all the modules deployed in the framework. This will, in turn, allow that any module can perform actions on one or more resources in response to a new state reached by a given resource. Actions executed in parallel from distributed environments.

4.2.1.1 Resources

Two main kinds of resources are considered: those derived from external sources such as (1) *documents* from textual files (e.g. a research paper), (2) *parts* from logical divisions of a *document* (e.g. rhetorical classes or sections), and (3) *domains* from sets of *documents* (e.g. a conference or journal), and those derived from processing the previous ones such as *annotations*.

To better illustrate this model, consider to explore the research papers published at the SIGGRAPH conference in 2016[5]. First, every paper will be materialized as a new *document* containing the full-text. Immediately after, the *document* will be automatically associated to several *parts*, each of them grouping sentences by rhetorical class (e.g. approach, background, challenge, future work and outcome) and by section (e.g abstract, introduction). Finally, a new *domain* will be created grouping all these *documents*. Different analysis will be performed extending the initial set of resources with more annotations at several representational levels: at *document level*, full-text based annotations are provided such as named-entities, compounds and descriptive tags. At *relational level*, connection between resources are found (e.g. semantic similarity-based relationships). And finally, at *domain level* annotations such as tags and summaries are composed describing the corpus of *documents*.

[5]http://s2016.siggraph.org

4.2.1.2 Event-based Paradigm

An event illustrates a performed action, i.e. a resource and its new state. It follows the Representational State Transfer (REST)Fielding and Taylor (2002) paradigm, but taking into account the state reached after an action, i.e *created*, *deleted* or *updated*. Thus, an event contains the resource type and the new state reached by a specific resource.

4.2.1.3 Linked Data Principles

Data in *librAIry* is individually addressable and linkable Turchi et al. (2012) following the Linked Data principles defined by T. Berners-Lee Bizer et al. (2009). Thus, resources (i.e. a *domain*, a *document*, a *part* or an *annotations*) have: (1) a URI as name, (2) a retrievable (or dereferenceable) HTTP URI so that it can be looked up, (3) a useful information provided by using standard notation (e.g. JavaScript Object Notation (JSON)) when it is looked up by URI, and (4) links to other URIs so that other resources can be discovered from it.

4.2.2 Framework Architecture

Following a publisher/subscriber approach, all the modules in the framework can publish and read events to notify and to be notified about the state of a resource. Therefore, the system flow is not unique and is not explicitly implemented, instead distributed and emergent flows can appear according to particular actions on resources.

4.2.2.1 Event-Bus

We use the Advanced Message Queuing Protocol (AMQP) as the messaging standard in *librAIry* to avoid any cross-platform problem and any dependency to the selected message broker. This protocol defines: *exchanges*, *queues*, *routing-keys* and *binding-keys* to communicate publishers and consumers.A message sent by a publisher to an exchange is tagged with a routing-key. Consumers matching that routing-key with the binding-key used to link the queue to that exchange will receive the message. In *librAIry* this key follows the structure: *resource.status*.Since a wildcard-based definition can be used to set the key, this paradigm allow modules both listening to individual type events

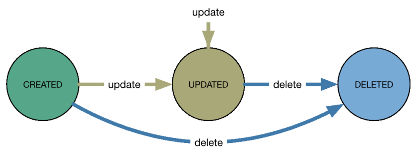

Figure 4.2: Resource states.

(e.g. ´domains.created´for new domains), or multiple type events (e.g. #.created for all new resources).

4.2.2.2 API

A HTTP-Rest Application Program Interface (API) was designed for interaction with end-users. Any external operation motivated by a user will be handled here. Some of them, usually those related to reading operations, will be completely managed by this module getting all the data from the internal storage. However, those operations implying a modification of the status of some resource (e.g. creation of a *document*), may be also performed by other modules listening for that type of event asynchronously. This module publishes to the following routing-keys: *domain.(created;updated;deleted)*, *document.(created;updated;deleted)*, *part.(created;updated;deleted)*, and *annotation.(created;updated;deleted)*

4.2.2.3 Storage

Multiple types of data can be handled in this ecosystem. Inspired in the Data Access Object (DAO) pattern, we have created a Unified Data Manager (UDM) providing access to any type of data used in the system. Three types of databases have been considered:

- **column-oriented database**: Focused on unique identified and/or *structured data*. This storage allow us searching key elements across resources.
- **document-oriented database**: Focused on indexing raw text. This storage allow us to execute advanced search operations over all the information gathered about a textual resource.
- **graph database**: Focused on relations. This storage allow us exploring resources through the relationships between them.

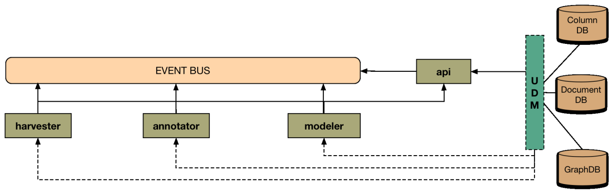

Figure 4.3: Modules.

4.2.2.4 Modules

The modules composing *librAIry* have been designed following the microservices architectural style. A module is a cohesive (i.e. it implements only functionalities strongly related to the concern that it is meant to model Dragoni et al. (2016)) and independent process working on the framework with a specific purpose. This purpose is defined by both the routing-key and the binding-key associated to the events handled by the module.

These are the main types of modules identified in *librAIry*:

- **Harvester**: creates system resources such as *documents*, *parts* and *domains*, from local or remote located textual files.
  - Listening for: nothing
  - Publishing to: *document.(created)*, *part.(created)*, *domain.(created;updated)*
- **Annotator**: retrieves named-entities, compounds, lemmas and other annotations resulting of Natural Language Processing (NLP) task execution from *documents* and *parts*.
  - Listening for: *document.(created;updated)*, *part.(created;updated)*
  - Publishing to: *annotation.(created;deleted)*
- **Modeler**: builds representational models from a given *domain*.
  - Listening for: *domain.(created;updated)*
  - Publishing to: *annotation.(created;deleted)*

4.3 Model-as-a-Service

...

4.4 Summary

In *librAIry*, existing algorithms and tools coming from different technologies can work collaboratively to process and analyze large collections of textual resources which has been successful applied to some real scenarios [6].

A new model definition based on the previously mentioned principle of maximizing information re-usability and minimize irrelevant data is being studied to create a more fine-grained resource design. New domains, in the sense of particular vocabularies or specific textual formats, are also being analyzed to be included into the system via specific harvesters or more precise annotators. Moreover, a template-based mechanism oriented to facilitate the integration of new tools and techniques into the system is being built to make easier to develop new modules as well as increasing the available modules at Docker-Hub.

[6] `http://drinventor.dia.fi.upm.es`

Chapter 5

Explainable Topic-based Associations

5.1 Topic Relevance

5.2 Topic-based Clustering

5.3 Summary

Chapter 6

Large-scale Comparisons of Topic Distributions

6.1 Document Similarity

6.2 Hashing Topic Distributions

6.3 Summary

Chapter 7

Cross-lingual Document Similarity

7.1 Synset-based Representational Space

7.2 Cross-lingual Models

7.3 Summary

Chapter 8

Evaluation

8.1 Evaluation Metrics

8.2 Text Representativeness

8.3 Large-scale Text Processing

librAIry has been used in some real scenarios such as a research-paper repository for the European project DrInventor [7], a support to decision makers for analyzing patents and public aids for the ICT sector, and also as a book recommender for an online content platform. This has allowed us to identify some weak and strong points of the framework and iterate over the architecture to come with the described solution.

The following modules have been developed[8]: (1) a ***general-purpose harvester*** which retrieves text and meta-information from PDF files in local or remote file-system; (2) a ***research paper-oriented harvester*** focused on collecting and processing more specific textual files (e.g. scientific papers) creating both *documents* and *parts* inferred from the rhetorical classes of the paper; (3) a ***Stanford CoreNLP-based Annotator*** which discovers named-entities, compounds and lemmas from *documents* and *parts*; (4) a ***Topic Modeler*** based on Latent Dirichlet Allocation (LDA) which creates probabilistic topic models for each *domain* in the framework. They are annotated with the set of topics (i.e. ranked list of words) discovered from the corpus, and both *documents* and *parts* of that domain are also annotated by the vector of probabilities to belong

[7] http://drinventor.eu

[8] https://github.com/librairy

to these topics. It uses the Spark implementation of the algorithm; and (5) a ***Word Embedding Modeler*** which creates a *word2vec* model from the *documents* contained in a *domain*.

Due to linear scalability and high performance features, Cassandra has been used to support the column-oriented storage functionality, Elasticsearch as document-oriented storage and Neo4j as graph-oriented storage.

All modules in *librAIry* have been packaged as Docker [9] containers and uploaded to Docker-Hub [10] to facilitate the installation of the system.

Maximizing information re-usability and minimize irrelevant data, becomes specially important when the system handles large collections of data (around million of documents). Fine-grained resource definitions have been key to achieve this, so modules execute actions only when really necessary. When a new *domain* is created, for instance, a new Topic Model is trained for that *domain* and is used to calculate the semantic similarity between the *documents* (and the *parts*) in that domain. If a new *document* (or *part*) is added to that *domain*, the model is trained again and the semantic similarities are re-calculated. However, this becomes unfeasible when the domain is frequently updated and it is composed by a large number of documents. One solution has been to define a new type of resource between domains and documents, models, that describes the representational state (e.g. topic model) of a collection of documents. Thus the model is only re-trained when a significant amount of *documents* are added to the sampling data set and not to the entire *domain*. This less transient model is used to calculate semantic similarities between the *document* collection (and *parts*) inside a *domain* in a more efficient way. Following this more precise execution of tasks, the routing-keys should include the URI of the implied resource into the definition, not only in the content of the message. It would allow modules listening to both the type of a resource or to a specific resource (or subsets, via regular expressions).

While the storage modules are always used to save/update/delete a resource, they are not always required from the end-user. The graph storage, for instance, makes sense when a path between two *documents* or *parts* is requested for a given *domain*. However, some *domains* are not intended to be explored by their linked resources. A

[9] https://www.docker.com
[10] https://hub.docker.com/u/librairy/

more fine/grained definition of resources will allow graph-storage being only used when necessary.

On the other hand, distributed execution of NLP tasks (not only in threads, but also in machines) has proved to be especially useful to handle large collection of *documents*. It requires less processing time than a monolithic solution (e.g. CoreNLP application) and it also provides a dynamic load balancing between modules.

8.4 Topic-based Clustering

8.5 Cross-lingual Similarity

8.6 Conclusions

Chapter 9

Experiments

9.1 Polypharmacy and Drug-drug Interactions

...

9.2 Corpus Viewer

...

9.3 ODS Classifier

...

9.4 Drugs4Covid

...

Chapter 10

Conclusions

10.1 Assumptions and Restrictions

...

10.2 Contributions

...

10.3 Impact

...

10.4 Limitations

...

10.5 Future Work

...

Bibliography

Badenes-Olmedo, C., Redondo-García, J., and Corcho, O. (2017a). An Initial Analysis of Topic-based Similarity among Scientific Documents Based on their Rhetorical Discourse Parts. In Garijo, D., van Hage, W., Kauppinen, T.and Kuhn, T., and Zhao, J., editors, *Proceedings of the First Workshop on Enabling Open Semantic Science co-located with 16th International Semantic Web Conference (ISWC)*, volume 1931 of *CEUR Workshop Proceedings*, pages 15–22. CEUR-WS.org. 13

Badenes-Olmedo, C., Redondo-Garcia, J., and Corcho, O. (2017b). Distributing Text Mining tasks with librAIry. In *17th ACM Symposium on Document Engineering (DocEng)*. 12

Badenes-Olmedo, C., Redondo-García, J., and Corcho, O. (2019a). Legal document retrieval across languages: topic hierarchies based on synsets. *arXiv e-prints*, page arXiv:1911.12637. 13, 14

Badenes-Olmedo, C., Redondo-García, J., and Corcho, O. (2019b). Scalable cross-lingual document similarity through language-specific concept hierarchies. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 147–153. 14

Badenes-Olmedo, C., Redondo-García, J., and Corcho, O. (2020). Large-Scale Semantic Exploration of Scientific Literature using Topic-based Hashing Algorithms. *Semantic Web*. 14

Badenes-Olmedo, C., Redondo-Garcia, J. L., and Corcho, O. (2017c). Efficient Clustering from Distributions over Topics. In *9th International Conference on Knowledge Capture (K-CAP)*, page 8. 13

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *International journal on Semantic Web and Information Systems*, 5(3):1–22. 20

Dragoni, N., Giallorenzo, S., Lafuente, A. L., Mazzara, M., Montesi, F., Mustafin, R., and Safina, L. (2016). Microservices: yesterday, today, and tomorrow. *CoRR*, abs/1606.0:1–17. 22

Fielding, R. T. and Taylor, R. N. (2002). Principled Design of the Modern Web Architecture. *ACM Transactions on Internet Technology*, 2(2):407–416. 20

Kosa, V., Chaves-Fraga, D., Naumenko, D., Yuschenko, E., Badenes-Olmedo, C., Ermolayev, V., and Birukou, A. (2018). Cross-Evaluation of Automated Term Extraction Tools by Measuring Terminological Saturation. In Bassiliades, N., Ermolayev, V., Fill, H., Yakovyna, V., Mayr, H. C., Nikitchenko, M., Zholtkevych, G., and Spivakovsky, A., editors, *Information and Communication Technologies in Education, Research, and Industrial Applications*, pages 135–163, Cham. Springer International Publishing.

Kosa, V., Chugunenko, H., Yuschenko, E., Badenes-Olmedo, C., Ermolayev, V., and Birukou, A. (2017). Semantic Saturation in Retrospective Text Document Collections. In Garijo, D., van Hage, W., Kauppinen, T.and Kuhn, T., and Zhao, J., editors, *Proceedings of the PhD Symposium at 13th International Conference on ICT in Education, Research, and Industrial Applications co-located with 13th International Conference on ICT in Education, Research, and Industrial Applications (ICTERI 2017)*, volume 1851 of *CEUR Workshop Proceedings*, pages 1–8. CEUR-WS.org.

López-Centeno, B., Badenes-Olmedo, C., Mataix-Sanjuan, Á., McAllister, K., Bellón, J. M., Gibbons, S., Balsalobre, P., Pérez-Latorre, L., Benedí, J., Marzolini, C., Aranguren-Oyarzábal, A., Khoo, S., Calvo-Alcántara, M. J., and Berenguer, J. (2019). Polypharmacy and Drug–Drug Interactions in People Living With Human Immunodeficiency Virus in the Region of Madrid, Spain: A Population-Based Study. *Clinical Infectious Diseases*.

Turchi, S., Ciofi, L., Paganelli, F., Pirri, F., and Giuli, D. (2012). Designing EPCIS through Linked Data and REST principles. *Software, Telecommunications and Computer Networks ({SoftCOM)}, 2012 20th International Conference on*, pages 1–6. 20