Analysis, evaluation and comparison of knowledge extraction tools in the Environmental and Health domain. A holistic approach

Anna Rovella
Università della Calabria, Italy
Alexander Murzaku
Saint Elizabeth University, USA
Eugenio Cesario
Università della Calabria, Italy
Martin Critelli
Università della Calabria, Italy
Armando Bartucci
Università di Macerata, Italy
Francesca M.C. Messiniti
Università della Calabria, Italy[1]

Abstract

Knowledge extraction in the Environment and Health domains is certainly an important asset for both scientific research and decision support. However, these strategic domains are characterized by a significant heterogeneity of structured and unstructured documents that do not allow a complete transfer of knowledge. Not infrequently, this critical point emerges during the implementation of research projects as an obstacle to the capitalization of information useful for the management of territories or in more recent times to the fight against the pandemic. This paper aims to achieve an analysis of the different forms of knowledge that characterize the scientific production in these specific fields trying to take a holistic approach to text management, tables and graphs through a multidisciplinary logic with the aim of

1 Authors contribution statement. Anna Rovella wrote and is responsible for the sections 1.0, 2.0, 3.0, 4.0 and 4.1. Francesca M.C. Messiniti carried out the experiments and is responsible for the statistical data of part 4.1. Alexander Murzaku wrote and is responsible for the section 4.2, Eugenio Cesario wrote and is responsible of section 4.3. Armando Bartucci wrote and performed the experiments of part 4.4 in collaboration with Francesca M.C. Messiniti. Martin Critelli wrote and performed the experiments of part 4.5 in collaboration with of Francesca M.C. Messiniti. All the authors contributed to the concept and design of the study, read and approved the final version of the article.

making the knowledge accessible through a geolocalized representation of sites at risk. A case study on a specific corpus of documents is provided.

1.0 Introduction

Environmental and Earth Observation domains provide a huge volume of heterogeneous research documents. This heterogeneity is also due to the natural intersection with other strategic domains, such as health and agriculture. For this reason, automatic knowledge acquisition and sharing becomes an important asset in this context.

For several years, the research community has worked to build open and shared knowledge bases. Despite these efforts, the need for knowledge extraction tools is still increasing. The relationship between Environmental research and Medical research has accentuated the need for rapid progress on disease-specific knowledge discovery. An example of this correlation is shown when assessing the impact of environmental pollution on the human body.

The large number of sources and the process of knowledge creation make information management a challenging process. In fact, without explicit sharing and effective communication, many data and research results are destined to a very limited use. Instead, an efficient process of knowledge sharing is useful in identifying agents and pathologies more quickly. Furthermore, this process allows a more accurate definition of high impact risk plans with positive effects on the prevention process.

Since the beginning of the pandemic, significant research has taken place aimed at finding new solutions to stop the spread of the contagion. The starting point of the research, which makes it possible, is the observation and collection of environmental data. In the Health and Environmental domain, the selection and definition of knowledge extraction tools, within a holistic vision, are essential for the efforts of researchers and decision-makers for creating and maintaining a non-hostile environment for humankind.

The purpose of this work is to analyse, evaluate, and compare tools for knowledge extraction from scientific literature specifying the described domain. In particular, the evaluation process aims to elaborate quantitative and qualitative data.

An integrated approach requires the identification of critical issues that documents, data and information contained within the Environmental and Health domain. This is a field characterized by structured and un-structured sources, textual documents defined on several levels and which include also objects such as tables and chart images.

The proposed approach aims to overcome some frequent problems in the informationextraction process from the reference literature. These include the extraction of content (metadata, keywords, entities, concepts, objects) but also the possibility of using experimental data often present in tables or images whose information is not immediately understandable and searchable.

To accomplish this goal, we start from the analysis of the performance of some knowledge extraction tools implemented for different and more specific purposes: extraction of metadata, keywords, terms and phrases, tables, charts, and images. The aim is the selection of a class of tools useful for the analysis of all dimensions of the content in the research documents. All the tools exploited in this work use machine learning or deep learning techniques along with different types of analysis and classification algorithms. The comparison and evaluation of all selected tools will be carried out on a specific set of domain documents. Special attention will be paid to the tools that show improved accuracy on the semantic level. The representation of the extracted data is the last step of our work. The purpose of this task is, for example, a possible use of the extracted data in decision support processes. The idea is to represent some extracted data geographical hotspot form and to return the images of tables or the charts in machine-readable form.

The rest of the paper is organized as follows. Section 2.0 presents work related to knowledge extraction from text, table, charts, metadata extraction and knowledge representation. In Section 3.0, we define the methodology of tests and evaluation followed by Section 4.0 in which the results of analysis, evaluation and comparison processes is discussed and an example of representation of geo-referenced data is presented. Finally, conclusions are drawn with notes on limitation and future research efforts are anticipated.

2.0 Related Work

Several techniques for automatic metadata extraction have been studied in the literature and various approaches have led to the implementation of many tools or frameworks. However, in the case of textual documents, structure is more complex. Most of the current tools are built to recognize and classify the basic structure of the input. The information extraction is defined as the identification of entities in the textual content within the document and the relationships of such entities with each other. In this domain, significant results have been achieved through the use of Machine Learning techniques (Liu et al. 2017).

The main limitation of Machine Learning techniques being the absence of data to train and finetune the classification models, we think that metada-

ta extraction tools could provide data to bootstrap this needed training corpus. In addition, we will evaluate the use of *BERT* (Bidirectional Encoder Representations from Transformers) techniques to enrich and better define this knowledge set. This would allow the improvement of the quality of the information extracted through the use of NLP technologies for parsing, tagging, and entity detection. We will apply and evaluate various NLP tools and packages such as Spacy for more precise and finer-grained analysis to research documents.

Table mining can be based on several approaches. They include table detection, functional analysis, structural analysis and semantic analysis. Each of these tasks can be accomplished through different techniques. Various frameworks for information extraction from tables based on multi-layers approaches with high precision scores have been proposed. For data extraction from chart images there are relatively novel Deep Learning approaches (Liu, Klabjan, and Bless 2019).

Moreover, the automatic extraction of geo-referenced data can play a fundamental role in enriching the knowledge model discovery task described above. For example, locations and places referred in the documents can enable the detection of spatial descriptive models, which could be valuable additional information for the Environmental and Health domain under analysis. This can be done by applying some spatial clustering algorithms for the discovery of geographic hotspots, aimed at detecting regions and areas where events of interest occur in with a higher density than other areas.

3.0 The method

The purpose of this work is to analyse, to evaluate, and to compare tools for knowledge extraction from scientific literature specifying the described domain. The results of this research can be subsequently used in various works and domains. One of these, for example, is the implementation of a platform of knowledge analysis and extraction, also in relation to the development of semantic models for the integration of heterogeneous knowledge.

First we define the corpus of documents for knowledge extraction. The selected corpus has already been validated by domain experts, in the European funded *e-shape* project. During that project, the experts more strongly highlighted the need to extract knowledge from images, tables and graphics. The corpus is a subset of scientific articles, extracted from PubMed (PCM database),[2] and concerning the impact of Mercury pollution on human

2 US National Library of Medicine, National Institutes of Health, "PubMed Central," last accessed September 28, 2021, https://www.ncbi.nlm.nih.gov/pmc/.

health. It consists of 85 articles on scientific journals in PDF files format. The subject of the articles is Mercury pollution diseases. The small size of the corpus is advantageous as it allows to manually check the results of the automatic extraction. So we were able to obtain a more precise evaluation of the performance of the tools.

We will focus on Machine Learning state-of-the-art solutions, that promise a more scalable solution and more rapid deployment ability.

First we select the tools on the basis of different criteria:

- tools with available open source code;
- tools of which we could verify the installation;
- tools preferably already applied to the Knowledge Base of PubMed.

The selected tools have been used to test their ability to extract knowledge on the chosen corpus. The results of the extraction have been measured and compared.

Some elements played an important role in the choice of tools. For example, for the extraction of metadata we have focused on an adaptive modular approach already tested on PubMed articles (Granitzer et al. 2012). Our purpose, in this case, was not just to check the performance of the tools. Rather, we were interested in understanding where and how to improve the semantic quality of the extracted content or how to adequately represent the information for immediate readability. Just as, in our holistic approach, it was important to test and evaluate the extraction of information from tables and charts and this we know is not a goal of metadata extraction.

The purpose of the linguistic analysis is limited to extracting shallow facts that can be repurposed later in the construction of a knowledge graph. The knowledge elements extracted were named entities, topics, and relations. Combining the extracted geographic named entities with topics provided the input necessary for the identification of geographic hotspots. Relations were detected through syntactic analysis of the text. Noun phrases and similar text chunks were encoded as nodes and verbs and prepositions as edges. Since the English language (all documents were in English) has a fixed order, the task of positioning nodes to the left or the right of the edge reflected the positions in the text itself. When syntactic labels were available, we modified the model to reflect the syntactic roles of the text chunks therefore having subject as left nodes, verbs as edges and objects as right nodes.

In the field of environmental data analysis, the detection of geographic hotspots is becoming a more and more popular task. In this work, we exploit a density-based clustering algorithm to perform a spatial partitioning of the area under investigation, where each cluster represents a dense region of toxicity due to heavy metal exposure. The density-based notion is a common approach for clustering, whose inspiring idea is that objects forming a dense

region should be grouped together into one cluster. In our implementation, this step is performed by applying DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Ester et al. 1996), a popular density-based clustering algorithm that finds clusters starting from the estimated density distribution of the considered data. We have chosen the DBSCAN algorithm because it has the ability to discover clusters with arbitrary shape such as linear, concave, oval, etc. and (in contrast to other clustering algorithms proposed in literature) it does not require the predetermination of the number of clusters to be discovered. Basically, the algorithm finds clusters with respect to the notion of density reachability among points: a point is directly density-reachable from another point if it is not farther away than a given distance (ϵ) (i.e., is part of its neighborhood) and if it is surrounded by sufficiently many points (minPts). In the considered context, a cluster corresponds to a heavy metal toxicity hotspot. Moreover, to capture the dynamic changing of clusters, we could compute the density of each data point by weighting it through a decay factor which gives less importance to historical information and more weight to recent data. Finally, DBSCAN requires the user to specify the radius of the neighborhood (i.e., ε) and the minimum number of objects it should have (i.e., *minPoints*), whose values affect size and density of the discovered clusters. Generally, an optimal setting of its parameters is complex to be achieved and requires specific techniques; nevertheless, such a topic is out of the scope of this paper.

Figure 1 shows a schematic representation of our research idea.

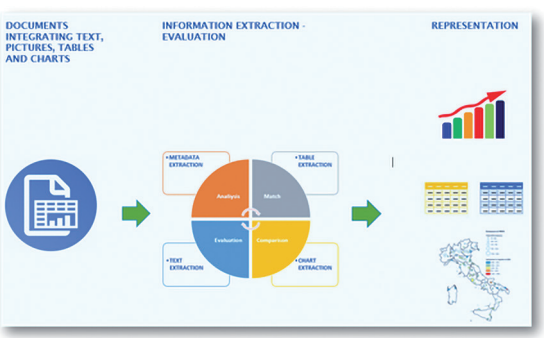


Figure 1. Schematic representation of work

4.0 Analysis, evaluation and representation of extracted knowledge

In this section we carried out the experiments to evaluate the extraction performance of the selected tools on the selected corpus. Each experiment describes processes running, metrics used and results.

4.1 Metadata extraction tool and experimental evaluation.

For metadata extraction we used Cermine (Tkaczyk et al. 2015), a framework created and trained to extract knowledge from PubMed. Cermine is an open source framework for extracting metadata and content from scientific article files in PDF format.

Its modular structure exploits supervised and unsupervised machine learning techniques (Support Vector Machines, K-means clustering and Conditional Random Fields). The System is a prototype developed in java for research purposes and its last update dates back to 2018.[3]

The output produced by Cermine is an xml file in the NLM JATS format.[4] The framework extracts, from documents, mostly Dublin Core metadata (title, author, affiliation, abstract, keywords, journal name, volume, bibliographic references, etc.).

Before describing our experiment, it is useful to recall how Cermine works for metadata extraction operating on the structure of the documents which is analyzed at different levels:

- the characters (dimensions and page coordinates) of the document are read and are identified;
- the different sections of the document are separated by geometric analysis of the pages (page segmentation);
- on the base of page segmentation, character recognition and heuristic structure analysis, the order of reading of the areas of the text is identified;
- then classification process associates metadata and different areas of the text;
- finally, the text is separated from the images, and is classified for the creation of two different output: a file for text and metadata (in NLM JATS format) and a directory for the images (png format files).

3 Due to the failure to update the software some framework tools (used for system training) are implemented with deprecated versions of Python 2.6.

4 American National Standard Developed by the National Information Standards Organization (ANSI/NISO). "Z39.96-2012 JATS: Journal Article Tag Suite". NISO, last updated July 26, 2013, https://groups.niso.org/apps/group_public/project/details.php?project_id=93.

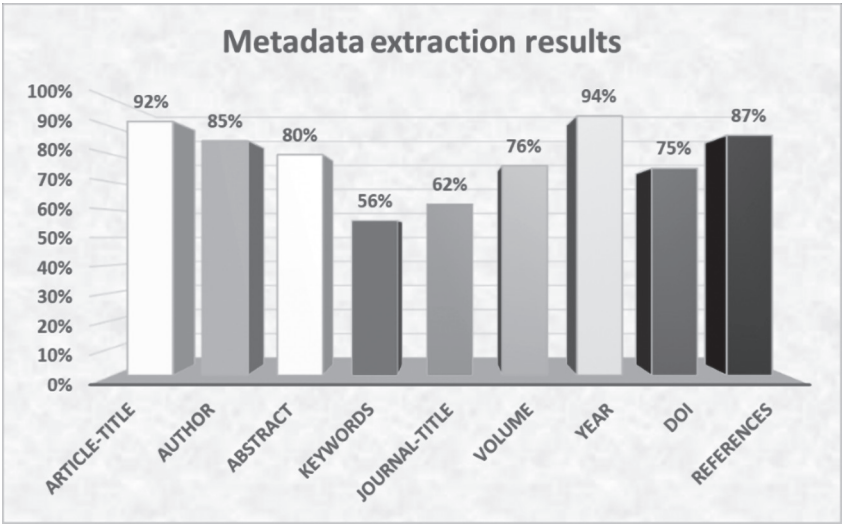


Figure 2. Metadata extraction results

For our experiment of metadata extraction and evaluation we worked according to the following tasks:

- corpus conversion and metadata extraction. Using OCR, we converted the corpus of pdf files into a searchable pdf format. Then we processed it with Cermine 1.13 standalone version. We used the original training set without any personalization;
- metadata extraction evaluation. After metadata extraction we proceeded with the evaluation of the output files by analyzing the quality of the results obtained. For this task we have implemented a specific tool. It is developed in Python with the aim to compare the Cermine output files with the NLM files downloaded from the PubMed Central subset. The tool measures comparison and analysis results by calculating recall and precision scores. Table and diagram form are the output of the tool for showing the metadata quality.

The tool checks the presence of the metadata tags in the file extracted. The values obtained show an extraction rate of more than 50% for all metadata and surprisingly the least extracted metadata are the keywords (56%) while, as can be seen, the year of publication is certainly a data that is always detected (Figure 2).

After this step, the algorithm used in the tool, analyzed the metadata values. This is done by comparing the string extracted by Cermine with the

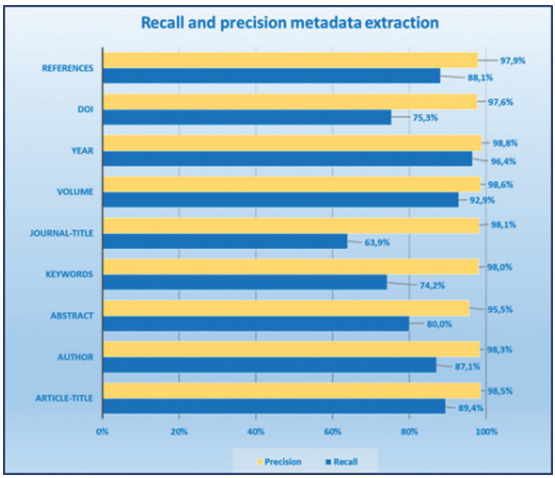


Figure 3. Calculation of recall and precision of metadata extraction

string stored in the NLM file. For the comparison of the strings and the determination of their similarity, the Levenshtein distance (or edit distance)[5] was used. The tool assigns a binary value in case of correct or incorrect extraction. The result of the extraction was considered null where the value strings were not complete.[6] These are mainly cases in which the layout models considered by Cermine without customizations are different from those analyzed. In these cases, we obtained a not optimized recognition of the areas of the text. Such situations would require a customization of the layout model that we do not take into consideration in this work. Finally, for each metadata extracted, recall and precision score are measured (see Figure 3). The chart shows that for some extracted metadata such as keywords and journal-title, Cermine obtains a result that is not optimized in quantitative terms (75%) despite the high quality of the information extracted.

From the calculation of the average of the extracted metadata values, we obtained an extraction evaluation with a precision value of 97.9%, recall 83% and error 16.9%.

5 "Levenshtein Algorithm," last accessed September 28, 2021, http://www.levenshtein.net/index.html.

6 These are mainly cases in which the layout models considered by Cermine without customizations are different from those analyzed. In these cases, we obtained a not optimized recognition of the areas of the text. Such situations would require a customization of the layout model that we do not take into consideration in this work.

A consideration comes from the extraction of the body metadata. This is present in a high percentage (96%), with a good quality of the information extracted, even when the article has a different layout from the models of the Cermine training set.

Cermine does not extract data from the images and is not able to recognize, with a good quality, data of the tables.

The evaluation of Cermine's extraction performance is overall positive. However, in order to improve the semantic quality of the extracted metadata, it can be assumed to apply NLP techniques to the analyzed texts. So, we proceeded with the next experiment.

4.2 Knowledge discovery from text

Knowledge discovery from text refers generally to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents (Tan, Mui, and Terrace 1999). The research and development of methods, that allow for fast and global analysis of textual data, created conditions for orienting not only decision-making but also the various research disciplines themselves. Given that a large part of scientific research is to understand previous research and to build on it, fast and accurate analysis of published work and knowledge discovery become the focus of attention for many institutions and regulatory organizations.

We are defining text as a general term for sequences of words. Text may be further structured into chapters, paragraphs, or sentences. For our purposes, the text unit that goes through linguistic analysis pipeline is the paragraph marked by the tag "<p>" in the XML output. However, this definition of text includes the concept of "word" which requires a further definition that leads to the concept of token and type. The distinction between a type and its tokens is an ontological one between a general sort of thing and its particular and concrete instances. Thus, 'do', 'does', 'done' and 'doing' are morphologically and graphically marked realizations of the same abstract word type 'do'(Gasparri and Marconi 2021). The process of identifying a token as type is also called lemmatization. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma (Manning, Raghavan, and Schütze 2008, 32). Indeed, in our process of knowledge discovery, the constituents used to identify objects and relations will be lemmas. The last sequence in need for a definition and fundamental for the process of knowledge extraction is the sentence. Sentences according to Quirk et al. (1991) are either simple or multiple. A simple sentence consists of a single

independent clause. Subject, verb, complement are constituents of sentences as well as of clauses within sentences. For identification purposes, a sentence is a sequence of words that has boundaries identified by punctuation marks. While question and exclamation points are relatively unambiguous markers of sentence boundary, periods are also present as abbreviation markers such as in etc., Mrs., or Inc. In general, sentence tokenization methods work by first deciding (based on rules or machine learning) whether a period is part of the word or is a sentence-boundary marker.

The text analysis pipeline includes the standard Python (v. 3.9) XML parser (xml.etree) for extracting text paragraphs from the XML files and spaCy (v. 3.1.3) (Honnibal et al. 2020) for tokenization/tagging, parsing/chunking, and named entity recognition (NER).

Text Pipeline

Through the tokenization process, we found that the corpus of 85 documents contains 165596 words/tokens, 16055 sentences, and 21135 unique words/types. Stop words, digits (when possible),[7] one-character long tokens, and punctuation signs are removed from these statistics.

Term frequencies

As expected, since the corpus is composed of articles researching mercury exposure, the list of the most frequent words includes [‘mercury’:3240, ‘exposure’:2359, ‘study’:2082, ‘Hg’:2059, ‘level’:1477, ‘blood’:907, ‘high’:866, ‘concentration’:759, ‘population’:747, ‘health’:665, ‘group’:661, ‘fish’:654, ‘biomarker’:636, ‘child’:628, …]. On a per-document basis, TF-IDF (Ramos 2003) is a better distinguisher of relevance as observed in this analyzed document where relevant words are sorted by TF-IDF score: [{doc:0, keywords: {tumour: 80.8, topsoil: 80.5, cancer: 67.5, mortality: 55.8, soil: 39.8, mainland: 34.6, metalloid: 32.3, town: 30.2, spain: 27.2, heavy: 26.3}}]

7 The presence of punctuation signs inside a sequence of digits (. and ,) is ambiguous on whether it is a decimal point or not depending on the locale.

Named Entity Recognition

The next step in linguistic analysis pipeline is the identification of named entities. Named entity recognition (NER) is the task of finding entities, such as people, locations, and organizations, in text.

The spaCy-recognized named entities are dominated by organizations/institutions (ORG: 6701), numerals (CARDINAL: 6516), dates/periods (DATE: 1940), countries/cities/states (GPE: 1920), and people names (PERSON: 1720). However, the use of a generic named entities recognizer causes a chemical formula (MeHg – methylmercury) to be recognized as a top GPE tag (China: 114, Japan: 62, US: 55, MeHg: 54, USA: 54, Spain: 52). Chaining a specialized NER package such as Chemlistem (Corbett and Boyle 2018) would allow distinguishing of domain specific terms. Also, the NER process exposes the need for coreference identification since in the top twenty GPE list we have US, USA, the United States, and U.S.

Topic detection

Topic detection is a useful mechanism for identifying various concepts embedded in a document, thus, allowing the user to navigate the collection of documents guided by topics. Topics are made up of relevant words, and they provide the user with an overview of the content of the individual documents as well as the document collection as a whole. Since in our sample of articles only 57% have a list of keywords (average 5.4 keywords per article), generating topic related lists of keywords becomes a useful corpus description instrument.

The packages we used are the *gensim* (Rehurek and Sojka 2011) package based on Latent Semantic Analysis (LSA) and the transformer based *BERTopic* (Grootendorst 2020).

Using gensim

In *gensim* every document is represented as a semantic vector. Using unsupervised machine learning algorithms, *gensim* allows for very fast processing and accurate results. The default number of topics is ten and each of them is illustrated by a cluster of ten words and the corresponding scores.

If we looke at the keyword list of the first article that had keywords, we can compare what is generated by *gensim* and what was entered in the publication:

Publication: ['amyotrophic', 'lateral', 'sclerosis', 'ALS', 'motor', 'neuron', 'disease', 'mercury', 'seafood', 'fish', 'consumption', 'dental', 'amalgam', 'filling', 'case-control', 'study', 'online', 'questionnaire', 'international', 'study']

Gensim: ['mercury', 'filling', 'seafood', 'ALS', 'occlusal', 'control', 'respondent', 'dental', 'exposure', 'current', 'proportion', 'online', 'factor', 'respondent', 'exposure', 'silver', 'eat', 'amalgam', 'questionnaire', 'consumption'][8]

The intersection is evident as highlighted by the underlined words.

Using BERTopic

This solution makes use of a sequence of techniques: it starts with the extraction of document embeddings using BERT (Devlin et al. 2018) and then reducing the dimensionality of embeddings to help the clustering process of the reduced embeddings. The output is a set of clusters of semantically similar documents. The final step is the extraction of representative keywords for each document cluster using Maximal Marginal Relevance (Carbonell and Goldstein 1998).

The keyword sets returned by *BERTopic* differ in the form they are organized from *gensim* even though semantically they cover the same meanings. Below is a list of the first 10 keyword groups out of 59.

['als', 'amyotrophic']

['mercury', 'methylmercury', 'methylamino']

['respondent', 'acknowledgment']

['filling', 'precipitate', 'cement']

['seafood', 'seafoods']

['control', 'motor', 'button']

['0111', '15', '1121', '046', '04', '007', '005', '001', '013']

['group', 'people', 'participant', 'human', 'somebody', 'individual', 'community', 'collect', 'committee', 'volunteer']

['dental', 'amalgam', 'tooth', 'mouth', 'bite', 'oral', 'chew']

The results of the location analysis are combined with the two topic extraction techniques allowing for a grouping of topics (such as those above) combined with the corresponding geographic locations such as ['Australia',

8 First topic keywords cluster augmented by keywords in the next topic clusters.

'Basel', 'Canada', 'Helsinki', 'Spain', 'Switzerland', 'USA']. This combination allows for a simple answer to the questions what and where.

Some issues with data

The topic modelling output, which is influenced by the relative frequencies of words (TF) as well as specificity of occurrences in the corpus (IDF), includes some peculiar word clusters. By analyzing them we conclude that repeated strings – and this is correct from the algorithmic point of view – are considered as relevant strings. These peculiar strings are generated by OCR errors (which, unfortunately are epxected), and by unexpected languages in the text. These strings affect the TF-IDF calculations of relevance and, therefore, distort keyword/topic detection results.

1. Number of languages included in a corpus

   For our experiment, we chose 85 English language articles; however, analysis shows a different story. Among the 507 unique characters present in the corpus there are:

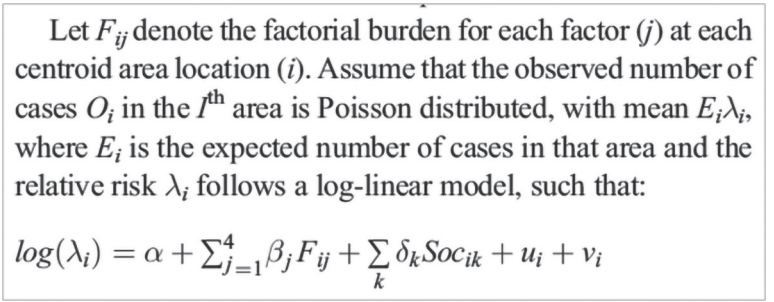
   i. Latin characters including accented characters more typical of romance languages: **ñ, è, î**
   ii. Greek characters: **μ, β, κ**
   iii. Arabic characters: **ت, ل, ن**
   iv. Cyrillic characters: **д, ы, л**
   v. Chinese/logographic characters: **考,地,女**

   The source of such strings is observed in the bibliography, location/person names, as well as in scientific formulas in the case of Greek.

2. Text extraction from PDF

   Mathematical and other scientific notation text segments generate a large amount of non-word strings as seen below (first the extracted text and second the screenshot of the PDF original text).

   Let Fij denote the factorial burden for each factor (j) at each centroid area location (i). Assume that the observed number of cases Oi in the Ith area is Poisson distributed, with mean Eiλi, where Ei is the expected number of cases in that area and the relative risk λi follows a log-linear model, such that:
   logðλiÞ ¼ α þ ∑4j ¼1β j Fij þ ∑ δk Socik þ ui þ vi


Let F_{ij} denote the factorial burden for each factor (j) at each centroid area location (i). Assume that the observed number of cases O_i in the I^{th} area is Poisson distributed, with mean $E_i\lambda_i$, where E_i is the expected number of cases in that area and the relative risk λ_i follows a log-linear model, such that:

$$log(\lambda_i) = \alpha + \sum_{j=1}^{4} \beta_j F_{ij} + \sum_{k} \delta_k Soc_{ik} + u_i + v_i$$

Figure 4. Screenshot of the PDF original text

An open-source software package like Tesseract[9] would allow the separation of scientific notation areas of the text from the rest of the text flow. This would significantly increase the quality of the extracted topics.

Relations

While extraction of topics and recognition of named entities give us a good view of who, what, and where – all of whom can be seen as nodes in a network – a knowledge network would also need a set of connection lines between these nodes. These lines or edges relate well to what in natural languages is expressed through verbs (and some prepositions). Starting with this assumption, we analyze our document corpus using automatic syntactic analysis.

Since the corpus contains documents in the English language, we take advantage of the order type of this language. Once the VERB at the root is identified, all the chunks on the left are considered to enter some relationship described by the verb in the chunks on the right therefore creating NODE-EDGE-NODE triples. Nodes (or chunks) such as *mercury* (758 occurrences), *exposure* (379), *the study* (91), *fish consumption* (89) relate to other nodes via the edges represented by verbs such as *show* (280 occurrences), *measure* (73), *increase* (53), *analyze* (37). For example, the verb *represent* is at the center of these relationships:

9 "Tesseract OCR," last accessed September 28, 2021, https://github.com/tesseract-ocr/tesseract#license.

'these maps' 'each symbol' 'these 75 countries' 'the cross-sectional studies' 'an example' 'nearly 50%' '48.6%' 'the data' 'terms' 'the number' 'individuals'	**represent**	'the average Hg concentration' 'an individual study' 'classes' 'graduated colours' 'the population subgroup' 'the reference' '4 countries' 'order' 'contribution' 'Republic' 'Korea' 'China' 'Japan' 'the United States'

Figure 5. Example of NODE-EDGE-NODE triples

As we are interested in the intersection of *what* and *where* we conclude with some data analysis focused on entities GPE and LOC identifying their adjacent dependent tokens (subject, verb, or object).

what	where
municipalities	Spain
people	USA
people	Australia
University	Montreal
have	Brazil
fell	Islands
came	States
live	Asia
representing	China
recyclers	India
population	USA

Figure 6. Examples of Entities

Notice that MeHg (discussed above) is found in the top occurrences in this corpus (China: 93, Europe: 58, MeHg: 36, Japan: 31, States: 25, Africa: 23, Spain, 20). The top of the *what*-column includes *children* (21 occurrences), *study* (15), *exposure* (13), *countries* (13), *population* (12), and *levels* (12). This approach allows for identifying both where certain issues are faced as well as what issues a certain location faces.

4.3 Discovery of Geographic Hotspots through density-based clustering, experimental Results

To evaluate the performance and the effectiveness of the proposed approach to discover geographic hotspots in a real-world case study, we carried out an extensive experimental analysis by executing different tests in a real scenario, i.e., a set of documents describing mercury toxicity cases occurred in the world.

As described above, geographic hotspots are detected by applying DBSCAN. As a first consideration before running the tests, in order to detect high quality city hotspots, it is necessary to tune the key parameters of the algorithm so as to improve performance results. DBSCAN takes in input two parameters, ε and *minPts*, which determines the size of the clusters, as they represent the minimum hotspot density required by an area to be part of a cluster. The bigger ε, the larger is the extension of the dense regions detected: this results in the discovery of large regions that actually are no longer dense. The smaller ε, the smaller the cluster sizes, resulting in a high number of dense hotspots detected that could be (because of their small sizes) not significant for the analysis. For what concerns *minPts*, it affects the density of the clusters, that is, the bigger (smaller) *minPts*, the lower (bigger) the average density of the detected clusters. We present here the results achieved by fixing $\varepsilon = 0.1$ and $minPts = 4$, which have been assessed through several experimental tests and best suits our application scenario and the considered dataset.

We performed pilot tests over the geographic data extracted from the documents and following the process described in Sections 4.1 and 4.2. The collected data and the achieved results obtained through our analysis are shown in Figure 7, 8 and 9.

In particular, Figure 7 shows the collected data (right side) and the discovered geographic hotspots (left side) about heavy metal issues discovered in Asia. Each hotspot is represented by a different color. Interestingly, this image shows how heavy metal issue events are clustered on the basis of a density criteria; for example, the algorithm detects several hotspots clearly recognizable through different colors: a large region (in red) in the top-right territory of China, along with several smaller areas (in green, blue and light-blue) on the left (China) and bottom (Japan) sides.

Figure 8 shows the collected data (right side) and the detected geographic hotspots (left side) discovered in Europe. There are clearly recognizable points covering Spain, Italy, United Kingdom, France and Denmark. Also in this case, the algorithm detects several hotspots identified by different colors: a large region (in light-blue) in the bottom-right territory of Spain, along with other areas diffused all over Europe.

Figure 7: Asia: mercury issues data and detected clusters

Figure 8. Europe: mercury issues data and detected clusters

Figure 9. North and South America: mercury issues data and detected clusters

Finally, Figure 9 shows the data and the detected geographic hotspots discovered in North and South America. In particular, several regions in the United States, Mexico and Peru are detected as hotspots of events, representing regions to be considered interesting for further analysis.

4.4 Table Extraction

There are several approaches for table extraction. Each of these approaches can be accomplished through at least two tasks:

- analysis of semi-structured documents based on mark-up language format (e.g., HTML or XML). Tags or coordinates of tables are computed to extract information from tables. However, the scientific articles rarely are already available in mark-up language format;
- pdf files conversion in semi-structured documents format based on HTML or XML languages. PDF is a widely used format in the scientific community for the production of articles. The PDF format does not provide information on the embedded physical layout.

However, it is not easy to convert unstructured documents into semi-structured ones. The main weakness of table extraction by converting PDF files depends on recognizing and understanding tables for automatic tools. This depends on:

- PDF format does not preserve the information related to the document's structure and the structure of tables. This information must be retrieved automatically from the way in which the text content is displayed. Furthermore, most automatic information extraction tools from scientific articles are developed on document layout analysis task based on machine learning algorithms. However, there is no single layout for scientific articles layout. For example, Kise (2014) identifies six kinds of document layout classes. Starting from this classification, Manhattan and Multi-Column Manhattan could be considered the most popular layout used in scientific articles but tables are collocated in different or in more text zones. This determines the dependence of the machine learning model on the single layout used with a consequent negative influence on the extraction results in the case of small layout differences;
- tables may have a different layout defined by the authors or based on the indications of the publishers. The absence of an international standard defining the rules for the creation of tables complicates the recognition and understanding of the model for machine learning-based extraction systems. For example, Luo et al. (2018) observe that the tables in the bio-

medical literature are often presented in a standard form of three-line tables and three-lines of information: *Caption*, *Header field* and *Data field*;
– the cells content can be heterogenous and can contain numbers or text or both. Furthermore, special characters (e.g. mathematical special character as ±) can be detected in an incorrect way.

Analyzing the tools available for table extraction represents an important task to define the state-of-art and propose possible future paths to improve information extraction from scientific articles and better knowledge dissemination.

The next paragraph proposes the analysis, evaluation and comparison tasks of knowledge extraction tools from tables in PDF document. During the experiments we made a comparison between the results obtained by extracting knowledge from tables using a special tool (Tabula) and using CERMINE, a metadata extraction framework. The results obtained from the extraction show the need for specific instruments. But let us describe the experiment.

Analysis and evaluation

The analysis and evaluation activity involved the search for tools for the automatic extraction of information from tables, the technical analysis, the evaluation of the advantages and disadvantages and finally the choice of tools. In the analysis step the main tools useful for this purpose have been identified. Most of them are not free, they do not support all kinds of operative systems, or they are not available because the link is not indicated in the articles. So, we chose Tabula, that has been implemented on a web browser in which the user can upload PDF file containing data table and browse the pages to manually or automatically detect tables by clicking and dragging to draw a box around the table. Tabula will extract data and it will allow to user to select final format (e. g. *.xls or *.csv). Manual detection improves the quality of the extraction but decreases the overall analysis time of the corpus, while the automatic detection decreases time, but the quality of the extraction is lower than manual detection. In this sense, Tabula could be useful to analyze a small corpus of documents (as in this case). For all these reasons, Tabula and CERMINE have been evaluated as the most suitable tools for the phase of extracting knowledge from the tables.

Comparison

The comparison of table mining has been based on the results obtained. We observed that the results can be classified on: “Totally extracted tables”,

"Partially extracted tables" and "Not extracted tables". A total of 232 tables have been extracted from the 85 articles of the corpus. Below, we report, an analysis of the results obtained with each software:

- CERMINE

The percentage results on the knowledge extraction from tables with CERMINE are classified in about 2% of the tables totally extracted, about 28% of the tables partially extracted and finally about 70% of the tables not extracted.

– In 4 cases the tables are completely extracted, and the results are immediately human readable;
– The errors of "Partially extracted tables" class are related to the extraction of a part of table (e.g. extraction of a single column or extraction of only the attributes of the columns) or not all tables in articles are extracted and also in this case they are partially extracted for the same reasons explained before. In this sense, "Partially extracted tables" cannot be read as good results;
– The errors of "Not extracted tables" class are related to the lack of extraction of tables present in the articles.

The reasons of negative results may depend on the strong dependence on the layout of scientific articles. In fact, CERMINE is trained on a Manhattan Layout model but some of the articles are based on the Multi-column Manhattan layout. Furthermore, CERMINE is unable to read tables if shown horizontally in articles and if placed on background texts (e. g. watermark).

The positive results regard the class of "Totally extracted tables". Although they do not contain specific characteristics required by NLM-JATS, all information is presented and could be used in the future for their conversion to NLM-JATS. Furthermore, an important result presented in most of "Table partially extracted" is extraction references. In many cases, CERMINE extracts the information on the references in NLM-JATS format by creating a link with the references section contained in the final XML.

- Tabula

The percentage results on the knowledge extraction from tables with Tabula are also classified in "Tables extracted totally", "Tables extracted partially" and "Tables not extracted" out of the total number of tables in the scientific articles analyzed equal to 232. In particular 87% of the tables were fully

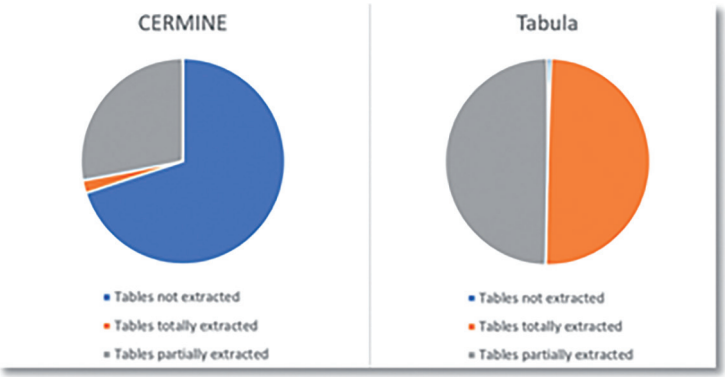


Figure 10. Percentages of "Tables totally extracted", "Tables partially extracted" and "Tables not extracted" using CERMINE and Tabula

extracted, about 12% of the tables were partially extracted and finally about 1% of the tables were not extracted. However, on 87% of the fully extracted tables, about 21% are readable and about 79% are unreadable.

The main limitation of Tabula is the way the information is read. In fact, Tabula reads each line in the tables from left to right and this causes the attributes to overlap between columns if they cover at least two lines of text. This negatively affects the result by determining the high percentage of unreadable tables. However, in this second case the percentage of "Tables extracted totally" tables are higher (caused by the manual detection) than "Tables extracted partially" and "Tables not extracted". It allowed to evaluate the quality of the extracted information. Furthermore, the manual detection makes the tool independent from the layout of the analyzed article but requires time to select each table in the text.

Final Consideration

The comparison considered the percentages obtained for each class of results ("Tables extracted totally", "Tables extracted partially" and "Tables not extracted") and the positive and negative cases of the extraction were used to evaluate the tools. At the end of this comparison, we can establish that Tabula shows a better quality of the information extracted from the corpus of scientific articles. In fact, manual recognition allows you to precisely identify the table in the article and make the tool independent from the layout. However, the use of Tabula is recommended for a limited number of documents. If not, we will spend a lot of time selecting all the tables.

4.5 Data extraction from charts

For charts extraction we analysed ChartOCR (Luo et al. 2021) and ChartReader.[10] ChartOCR is a Deep Learning based framework developed by Luo et al. (2021) for Ubuntu systems and is able to perform data extraction making a Data Table as output. This framework is implemented using CNNs architectures with a Microsoft OCR API to extract text from the image. Since this architecture is very complex, the framework requires a remarkable GPU computing power -in the original experimentation 4 Tesla P100 GPUs were employed. The framework first extracts common information in this case the chart type recognition is performed through the detection of key points. The next phase is the extraction of data range. The data range is calculated in order to read the numerical values inside the graph and, in the final phase, it allows the extraction of data according to the specific type of graph. The last task of the framework is implemented using CNNs architectures with a Microsoft OCR API with the aim to extract text from the image. Since the complex architecture, the framework requires a remarkable GPU computing power.[11] Due to the high GPU required in this work we decided to use another framework, ChartReader, which is characterized by a lighter computer architecture. It is developed by C. Rane and is available on the GitHub page of the author. ChartReader is composed by several modules useful for different purposes:

1. extracting DOI from the PDF articles;
2. recognizing type of chart and axis labels using two different CNNs architectures: VGG-19 & EfficientNetB3;
3. extracting text using AWS API from plots;
4. extracting data.

A last code allows to collect all the extracted information, saved in JSON, inside a single CSV file.

In our experimentation, we use ChartReader as a test on chart images set extracted from the corpus in order to prove their efficiency. The model chosen for testing ChartReader was VGG-19 because its inferior time requirement per inference compared to that required by EfficientNetB3.[12] For the testing phase we used the GPU available on Google Colab. As output we

10 Chinmayee Rane, "ChartReader," last accessed September 28, 2021, https://github.com/Cvrane/ChartReader.

11 In the original experimentation Luo et al. (2021) declare having employed 4 Tesla P100 GPUs.

12 For further details visit "Keras," last accessed September 28, 2021, https://keras.io/api/applications/.

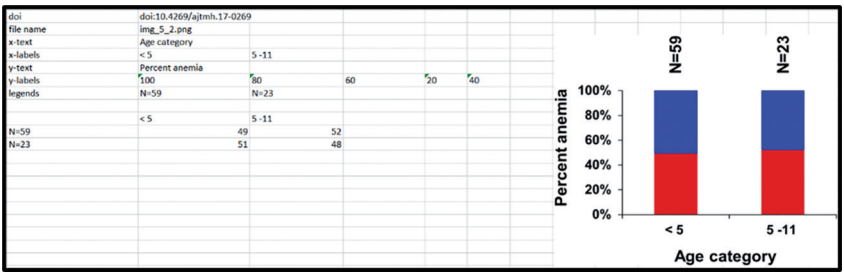


Figure 11: Data & Text extraction using ChartReader

obtained the CSV file which contains all information extracted through the modules described above.

Figure 11 reports a sample of the performed extraction.

5.0 Conclusion and future work

In this paper we proposed a holistic approach to the extraction of knowledge and the representation of information in the environment and health domains. The approach was also exemplified by an experiment conducted on a corpus of 85 scientific papers from PUBMED. The experiment was conducted with a multidisciplinary logic that allowed us, through the application of tools and predictive algorithms, automatic extraction of metadata, text analysis for automatic extraction of content (terms, objects, subjects, entities, relationships), the automatic extraction of data and information from tables and charts, and finally the geolocalized representation of sites at risk. For the experiment we used non customized opensource applications. Although some of the technologies we have used require further optimisation efforts, our approach has shown significant results not achievable on average through one-way approaches. The holistic approach has revealed interesting potential for positive repercussions in the context of research as well as in support of decision-making. The future developments of our research will mainly concern the customization of the tools used and their targeted training also in a logic of integration for the construction of an innovative framework of knowledge extraction.

References

American National Standard Developed by the National Information Standards Organization (ANSI/NISO). "Z39.96-2012 JATS: Journal Article Tag Suite". NISO.

Last updated July 26, 2013. https://groups.niso.org/apps/group_public/project/details.php?project_id=93.

Carbonell, Jaime, and Jade Goldstein. 1998. "The use of MMR, diversity-based reranking for reordering documents and producing summaries." In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval August 1998*, 335–6. https://doi.org/10.1145/290941.291025.

Corbett, Peter, and John Boyle. 2018. "Chemlistem: chemical named entity recognition using recurrent neural networks." *Journal of Cheminformatics* 10, no. 59, Springer Nature. https://doi.org/10.1186/s13321-018-0313-8.

Devlin, Jacob, Ming-Wei Chng, Kenton Lee, and Kristina Tautanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Cornell University. http://arxiv.org/abs/1810.04805.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining 2-4 August 1996 Portland, Oregon*, edited by Evangelos Simoudis, Jiawei Han, and Usama Fayyad, 226-31. ISBN 978-1-57735-004-0.

Gasparri, Luca, and Diego Marconi, "Word Meaning", The Stanford Encyclopedia of Philosophy, (Spring 2021 Edition), Edward N. Zalta (ed.), August 9, 2019, https://plato.stanford.edu/archives/spr2021/entries/word-meaning/.

Granitzer, Michael, Maya Hristakeva, Kris Jack, and Robert Knight. 2012. "A comparison of metadata extraction techniques for crowdsourced bibliographic metadata management." In *Proceedings of the 27th Annual ACM Symposium on Applied Computing, (ACM 2012) March 2012 Trento, Italy*, 962-4. https://doi.org/10.1145/2245276.2245462.

Grootendorst, Maarten. 2020. "BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics (Version v0.7.0)." https://doi:10.5281/zenodo.4381785.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. "Industrial-strength Natural Language Processing in Python." https://10.5281/zenodo.1212303.

"Keras." Last accessed September 28, 2021. https://keras.io/api/applications/.

Kise, Koichi. 2014. "Page Segmentation techniques in document analysis." In *Handbook of Document Image Processing and Recognition*, edited by D. Doermann and K. Tombre, 134-75. London:Springer. https://doi.org/10.1007/978-0-85729-859-1_5.

"Levenshtein Algorithm." Last accessed September 28, 2021. http://www.levenshtein.net/index.html.

Liu, Runtao, Liangcai Gao, Dong An, Zhuoren Jiang, and Zhi Tang. 2017. "Automatic Document Metadata Extraction based on Deep Networks." *Natural Language Processing and Chinese Computing, LNCS* 10619, edited by Huang, Xuanjing, Jing Jiang, Dongyan Zhao,Yansong Feng and Yu Hong. Springer, 305-17. https://doi.org/10.1007/978-3-319-73618-1_26.

Liu, Xiaouyi, Diego Klabjan, and Patrick NBless. 2019. "Data Extraction from Charts via Single Deep Neural Network." arXiv:1906.11906v1.

Luo, Daipeng, Jing Peng, and Yuhua Fu. 2018. "Biotable: A Tool to Extract Semantic Structure of Ta-ble in Biology Literature." In *ICBRA '18: Proceeding of the 2018 5th International Conference on Bioinformatics Research and Applications (ACM 2018) 27-29 December 2018 Hong Kong, Hong Kong,* New York: Association for Computing Machinery, 29-33. https://doi.org/10.1145/3309129.3309139.

Luo, Junyu, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. 2021. "ChartOCR: Data extraction From Charts Images via a Deep Hybrid Framework." In *Proceedings of the IEEE/CVF Winter Con-ference on Applications of Computer Vision 3-8 January 2021 Waikoloa, HI, USA*, 1916-1924. https://doi.org/10.1109/WACV48630.2021.00196.

Manning, D. Christopher, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Quirk, Randolph, Sidney Greebaum, Geoffrey Leech, and Jan Svartvik. 1991. *A Comprehensive Grammar of the English Language*. London: Longman.

Ramos, Juan. 2003. "Using tf-idf to determine word relevance in document queries." In *Proceedings of the first instructional conference on machine learning* 242, no. 1: 29-48.

Rane, Chinmayee, "ChartReader". Last accessed September 28, 2021. https://github.com/Cvrane/ChartReader.

Rehurek, Radim, and Petr Sojka. 2011. "Gensim–python framework for vector space modelling." *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3, no. 2.

Tan, Ah-hwee. 1999. "Text mining: The state of the art and the challenges." *In Proceedings of the pakdd 1999, workshop on knowledge discovery from advanced databases* 8, 65-70.

"Tesseract OCR." Last accessed September 28, 2021. https://github.com/tesseract-ocr/tesseract#license.

Tkaczyk, Dominika, Pawel Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Lukasz Boli-kowski. 2015. "CERMINE: automatic extraction of structured metadata from scientific literature." *International Journal on Document Analysis and Recognition* 18, no. 4: 317-35. https://doi.org/10.1007/s10032-015-0249-8.

US National Library of Medicine, National Institutes of Health, "PubMed Central." Last accessed September 28, 2021. https://www.ncbi.nlm.nih.gov/pmc/.