T2Know: An Advance Scientific-Tecnical Text Analysis Platform for Trend and Knowledge Extraction Using NLP Techniques

Rafael Muñoz, Yoan Gutiérrez and Andrés Montoyo

University of Alicante, Spain. Crta. San Vicentte del Raspeig s/n, Alicante, Spain

Abstract

The project T2Know presents the use of natural language processing technologies for the creation of a semantic platform of scientific documents via knowledge graphs. This knowledge graph will link relevant parts of each document with those of other documents in such a way that trend analysis and recommendations can be achieved. The goals addressed within the scope of this project include entity recognizers development, profile definition and documents linkage through the use of transformers technologies. As a result, the relevant parts of the documents to be extracted are related not only to the title and affiliation of the authors, but also to article topics such as references, which are also considered relevant parts of the scientific article.

Keywords

Semantics, semantic document profile, entity recognition, language models, trasnformers,

1. Introduction

Health research organizations have always been an exceptional environment for identifying specific needs and generating new ideas that lead to innovative processes, products and services that improve health outcomes and the sustainability of health systems. Additionally, these organizations also provide key information on the environment and market trends in their scientific areas of interest.

Nevertheless, despite having technological surveillance (in many cases) or competitive intelligence systems that enable them to configure customized alerts or the option of performing specialized information retrieval searches, they still lack solutions that support the systematic analysis of the large volume of information to retrieve the desired information. More specifically, according to industry figures, medical professionals can spend up to 20 percent [1] of their working day conducting information searches to support their daily activities.

In this context, in order to provide greater value, both to society and to the people who make up the structure of health research institutes, we propose the implementation of a platform for the advanced analysis of large volumes of textual data in the form of digital documents of a scientific-technical nature. This facilitates systematic analysis, environment assessment and tje proposition of plausible future scenarios. This project significantly supports the transition from a traditional model of medicine, known as reactive or curative medicine, in which patients go to the doctor and are treated, to a medicine that is not satisfied with curing, but seeks to prevent, improve the quality of life, adapt to the individual, predict the evolution of the disease and put the patient at the center.

In this process of change towards a new paradigm, as in the case of 5P medicine — which stands for more preventive, participatory, personalized, predictive and population-based medicine—, HLTs, and tools such as the one proposed in this project, play a fundamental role owing to the great potential offered by health data collection and the analysis of large collections of health data.

A convenient RDI planning, adapted to both the organization and its environment, will guarantee the optimization and investment in technological developments that meet society's demands. Thus, this would ensure effectively translating these aspects into clinical practice through the productive ecosystem. Trend identification in research that consequently creates new markets will flourish the development of companies which respond to these new business niches. This will translate into a significant improvement in patients' quality of life and in the generation of wealth and well-being in society

1.1. Project objectives

The main objective of this project is the research and development of T2KNOW, an advanced text analysis

SEPLN-PD 2023: Annual Conference of the Spanish Association for Natural Language Processing 2023: Projects and System Demonstrations

rafael@dlsi.ua.es (R. Muñoz); ygutierrez@dlsi.ua.es (Y. Gutiérrez); montoyo@dlsi.ua.es (A. Montoyo)
0000-0001-8127-9012 (R. Muñoz); 0000-0002-4052-7427 (Y. Gutiérrez); 0000-0002-3076-0890 (A. Montoyo)

[1] https://www.consalud.es/profesionales/los-medicos-podrian-sufrir-infoxicacion_13126_102.html

platform based on Natural Language Processing (NLP) technologies, is the extraction and representation of semantic profiles of digital entities and the identification of research trends from the automatic analysis of scientific-technical documents. Starting from this general objective, the project has the following specific objectives:

- To design and develop a flexible, scalable and robust technological architecture for the management and processing of large volumes of unstructured data (text) as a necessary basis for advanced analysis.
- To research and develop advanced text analysis algorithms, using PLN techniques, that allow knowledge extraction and semantic exploration of content for the detection of research trends.
- To develop data visualization technologies to discover and graphically represent the evolution of research lines, topics and emerging technologies that allow the identification of research trends.
- To design and execute a pilot test to validate the technologies developed in a key area such as healthcare, with the creation of specific corpus of scientific publications.

2. State of the art

The process of knowledge discovery from natural language can be seen as a flow composed of several stages: from the initial text to a final relevant semantic knowledge representation in the form of an ontology. The first step consists in the manual or semi-automatic construction of annotated linguistic resources. This requires choosing or defining an annotation scheme that is conducive to the domain of interest. From an annotated corpus, machine learning algorithms are trained to apply the same annotation scheme to large volumes of text. Subsequently, all automatically discovered entities and relationships are grouped into a semantic graph. At this point, it is possible to perform post-processing tasks to eliminate redundancies, combine similar entities, or detect inconsistencies. Finally, a unified semantic structure is obtained, which can be presented in the ontology format, where the relevant knowledge that was implicit in the original text is represented.

To extract relevant knowledge from natural language text, PLN techniques have been introduced in systems such as ISODLE [1].

The use of natural language features can be used to build rule-based systems, such as the proposed OntoLT [2], which extracts concepts and relationships through a mapping from linguistic classes to ontology classes. An alternative approach is the use statistical or probabilistic models, exemplified by systems such as LEILA [3] or Text2Onto [4]. Another example is KnowItAll [5], which introduces a pointwise mutual information (PMI) metric to select relevant instances. Once instances of entities and relationships are extracted from the text, a natural question is whether more abstract knowledge can be inferred from these examples. Systems that address this problem often use unsupervised techniques to attempt to discover inherent structures. Two relevant examples of this approach are OntoGain [6] and ASIUM [7], which attempt to automatically construct a hierarchy of concepts using clustering techniques. Even though most of the aforementioned systems generally focus on one iteration of the extraction process, more recent approaches, such as NELL [8], attempt to continuously learn from a stream of web data and increase over time both the quantity and quality of the knowledge discovered.

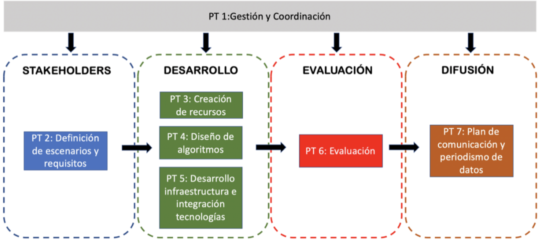
PT 1:Gestión y Coordinación
STAKEHOLDERS
DESARROLLO
EVALUACIÓN
DIFUSIÓN
PT 2: Definición de escenarios y requisitos
PT 3: Creación de recursos
PT 4: Diseño de algoritmos
PT 5: Desarrollo infraestructura e integración tecnologías
PT 6: Evaluación
PT 7: Plan de comunicación y periodismo de datos

Figure 1: Work packages organization.

As we have seen, in the systems mentioned above, in order to extract knowledge from textual sources, it is necessary to contemplate the use and development of techniques for the semantic representation of knowledge, its storage and computational processing, and its metrics for the evaluation of its quality.

3. Human language technologies

The T2Know project is a consortium involving several entities such as ISABIAL (Institute of Health and Biomedical Research of Alicante), the company DIFUSION S.L. and the University of Alicante. In addition, stakholders such as AIMPLAS (Instituto Tecnológico del Plástico) and ITI (Centro Tecnológico de Investigación, Desarrollo e Innovación TIC) were invited to be part of the team, since they will contribute with expertise in areas related to the health field such as plastics or technology. Depending on the role played by each entity, these appear in different modules, as shown in Figure 1.

The focuses on three use cases or application areas such as health, plastics and technology. The focuses on three use cases or application areas such as health, plastics and technology. For these, taxonomies have been identified, such as the one in the table 1 for health.

Table 1
Uses case

Scientific and technical support	Topics y subtopics	Example of associated Taxonomy
Health	Biomarkers in neurodegenerative diseases: Huntington's, Multiple Sclerosis and Alzheimer's.	• Huntington – Epigenetics – Biomarker – Transcription – Next Generation Sequencing – Etc. • Multiple Sclerosis – Myelin – Trained immunity – Olygodendrocyte – Epigenetics – Ageing – Inflammation – Etc. • Alzheimer – Extracellular vesicles – Exosomes – Etc.

3.1. Document profile representation

This task is responsible for semantically representing documents based on a series of characteristics previously extracted from them. This representation is governed by the scheme defined in the figure **??**, which will allow to characterize the documents and in turn the entities and elements included in them, so that, with the use of semantic exploration techniques, it will be possible to recover not only documents but also their metadata as digital entities, with their respective profiles.

For instance, a given document may have a title, authors, an abstract, different topics and entities and also citations to other documents. In turn, authors may have an associated email, affiliation to one or several institutions, which may have an associated country. As can be seen, there are several characteristics that, when analyzed at a deeper level of detail, reveal the benefit of representing everything that can be characterized, since it enriches the quality of the metadata and thus offers greater opportunities and points of view to explore the documents. Therefore, once documents are semantically represented, it will be possible to query not only documents, but also metadata under multiple non-conventional criteria such as, for example, Countries, institutions or authors most active in a given topic.

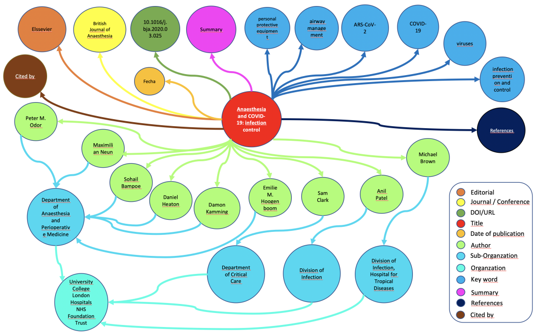
Elssevier
British Journal of Anaesthesia
10.1016/j.bja.2020.03.025
Summary
personal protective equipment
airway management
ARS-CoV-2
COVID-19
viruses
infection prevention and control
Cited by
Fecha
Anaesthesia and COVID-19: infection control
References
Peter M. Odor
Maximilian Neun
Sohail Bampoe
Daniel Heaton
Damon Kamming
Emilie M. Hoogenboom
Sam Clark
Anil Patel
Michael Brown
Department of Anaesthesia and Perioperative Medicine
Department of Critical Care
Division of Infection
Division of Infection, Hospital for Tropical Diseases
University College London Hospitals NHS Foundation Trust
Editorial
Journal / Conference
DOI/URL
Title
Date of publication
Author
Sub-Organzation
Organzation
Key word
Summary
References
Cited by

Figure 2: Semantic profile of documents.

3.2. Development of technologies for semantic information extraction from documents

Semantic information extraction is a task in which, by relying mainly on a set of PLN sensors, pieces of information with semantic connotations are extracted from the textual content. These sensors must be able to identify, classify and extract all the necessary information to create a knowledge base. In this task, NERC domain and text classification sensors are developed, allowing the application of information and terminology extraction processes. In particular, the following phases or stages will be followed:

- Terminology extraction. Term extraction based on statistical and linguistic algorithms.
- Domain entity recognition. Development of entity recognizers specialized in each type of concept.
- Information extraction through PLN tools, e.g., sensors of attributes defined in the document profile such as date, author, language, summary, topics, entities, among others.

To be able to identify, at a general level, different categories to which documents may belong, and to be able to classify within them the different types of domain entities, it is necessary to reuse and develop different sensors based on natural language processing technologies. Therefore, machine learning models will be developed, adapted and reused for this purpose.

With these sensors it will be possible to classify documents according to different topics, as well as to identify and categorize different types of entities present in the contents, in order to guarantee an advanced exploration of the knowledge involved. To make possible the development of these sensors based in part on machine learning, we will start from the corpus annotated by domain experts that will be used to train the PLN models.

3.3. Profile linking

During this task, processes will be developed to link the document profiles through semantic relationships, as shown in the figure 3.3. That is, once the documents are semantically represented, it is necessary to link them to each other, and in this way they will serve as a connection point between other digital entities such as, for example, authors, subjects, etc. Not only will the documents serve as a link between digital entities, but also the common characteristics found between documents already semantically represented.

This linking of profile characteristics enables the inferring and discovering of new information otherwise harder to identify at first sight. For instance, authors or institutions that coincide on a particular topic or have written about a particular technology.

Profile linking presents the problem of semantic ambiguity [9] and ontology mapping [10] when it is proposed to automatically link profiles without running the risk of making mistakes. This issue can be dealt with in a semi-supervised way by a human and using semantic techniques to discern what would be the appropriate link between profiles, either of documents or of any digital entity involved (ie. authors, institutions, countries, etc.).

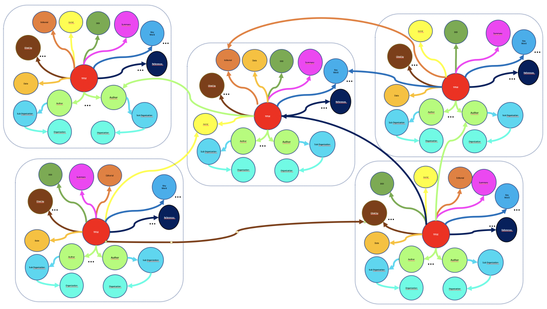

Figure 3: Document semantic profile linking

3.4. Data analytics and trends

In this task, tools will be developed to detect the greatest amount of statistical information related to the document profiles by temporal fractions. This will allow the trend identification and topic evolution related to research in the sectors involved in the project. For this purpose, automatic learning techniques will be resorted to, such as Time Series, as well as the potential offered by their visualization. This visualization allows a two-fold result: (1) the generation of new knowledge, and (2) the presentation of the temporal and statistical evolution of the pieces of information involved in the identified ontology.

The processes involved in the ontology life cycle, namely, the ontology creation, management, analysis and reuse, entail workflows formed by several activities. These have been defined taking into consideration the main methodologies for the development of ontology models. Additionally, it is also necessary that these activities are supported by mechanisms and tools that allow their efficient development. These mechanisms are mainly robust visualization techniques and provided with an interaction that allows the user, through its capacity, to develop abstraction, conception, understanding, representation and learning of knowledge. One of the most important aspects to take into account in this task is semantic exploration and recommendation. Given the existence of a semantic database, it is necessary to develop mechanisms for the exploration of the semantic network, supported by SPARQL(https://skos.um.es/TR/rdf-sparql-query/) or Cypher (https://neo4j.com/developer/cypher/) queries, of document profiles and other digital entities. These mechanisms will allow to retrieve not only documents through metadata filters, but also to make aggregate queries to discover statistical trends, and to make recommendations of profiles (e.g., documents, authors, institutions, topics, named entities, etc.) through the semantic links that interconnect the network.

4. Conclusions

Currently, the project is in an initial stage focused on the capturing of both technical and functional requirements. In addition, the KPIs, of key importance for identification, have been identified and the system development and evaluation will begin shortly.

Acknowledgments

This project is funded by the Valencian Agency for Innovation through the project INNEST/2022/24, partially funded by the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) through the following projects NL4DISMIS: TLHs for an Equal and Accessible Inclusive Society (CIPROM/2021/021) and T2Know: Platform for advanced analysis of scientific-technical texts to extract trends and knowledge through NLP techniques.(Innest/2022/24). Moreover, it was backed by the work of two COST Actions: CA19134 - "Distributed Knowledge Graphs" and CA19142 - "Leading Platform for European Citizens, Industries, Academia, and Policymakers in Media Accessibility".

References

[1] N. Weber, P. Buitelaar, Web-based ontology learning with isolde, in: Proc. of the Workshop on Web Content Mining with Human Language at the International Semantic Web Conference, Athens GA, USA, volume 11, 2006.

[2] P. Buitelaar, M. Sintek, Ontolt version 1.0: Middleware for ontology extraction from text, in: Proc. of the Demo Session at the International Semantic Web Conference, 2004.

[3] F. M. Suchanek, G. Ifrim, G. Weikum, Leila: Learning to extract information by linguistic analysis, in: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, 2006, pp. 18–25.

[4] P. Cimiano, J. Völker, text2onto, in: International Conference on Application of Natural Language to Information Systems, Springer, 2005, pp. 227–238.

[5] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates, Web-scale information extraction in knowitall:(preliminary results), in: Proceedings of the 13th international conference on World Wide Web, ACM, 2004, pp. 100–110.

[6] E. Drymonas, K. Zervanou, E. G. Petrakis, Unsupervised ontology acquisition from plain texts: the OntoGain system, in: International Conference on Application of Natural Language to Information Systems, Springer, 2010, pp. 277–287.

[7] D. Faure, T. Poibeau, First experiments of using semantic knowledge learned by asium for information extraction task using intex, in: Proceedings of the ECAI workshop on Ontology Learning, 2000.

[8] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, et al., Never-ending learning, Communications of the ACM 61 (2018) 103–115.

[9] Y. Gutiérrez, S. Vázquez, A. Montoyo, Spreading semantic information by word sense disambiguation, Knowledge-Based Systems 132 (2017) 47–61.

[10] Y. Gutierrez, D. Tomas, I. Moreno, Developing an ontology schema for enriching and linking digital media assets, Future Generation Computer Systems 101 (2019) 381–397.