T2Know: Analysis and Trend Platform using the Knowledge Extracted from Scientific Texts

Rafael Muñoz University of Alicante
Manuel Palomar University of Alicante
Yoan Gutíerrez University of Alicante
{rafael|mpalomar|ygutierrez}@dlsi.ua.es
Mar Bonora University of Alicante
mar.bonora@ua.es

Abstract

The T2Know project explores the application of natural language processing technologies to build a semantic platform for scientific documents using knowledge graphs. These graphs will interconnect meaningful sections from different documents, enabling both trend analysis and the generation of informed recommendations. The project's objectives include the development of entity recognition systems, the definition of user and document profiles, and the linking of documents through transformer-based technologies. Consequently, the extracted relevant content will go beyond standard metadata such as titles and author affiliations, extending also to other key sections of scientific articles, including references, which are treated as integral components of the knowledge representation.

1 Introduction

Health research organizations have long stood out as ideal environments for detecting specific needs and fostering the emergence of innovative ideas. These ideas often give rise to new processes, products, and services that not only enhance health outcomes but also contribute to the sustainability of healthcare systems. In addition, such organizations offer valuable insights into their scientific fields of interest by monitoring environmental and market trends.

Despite the availability of technological surveillance tools and, in many cases, competitive intelligence systems that enable the configuration of customized alerts and advanced information retrieval, these organizations still face significant challenges. Specifically, they lack robust solutions that allow for the systematic analysis of the vast amount of information required to extract relevant and useful knowledge. Industry data reveals that medical professionals can spend up to 20% of their working hours searching for information to support their daily tasks.

In light of this, and with the aim of delivering greater value to both society and the professionals within health research institutes, we propose the development of a platform capable of advanced analysis of large volumes of scientific and technical textual data. This would facilitate systematic information processing, environmental analysis, and the formulation of plausible future scenarios. Such a tool would greatly support the ongoing transformation from traditional, reactive healthcare—where treatment follows the onset of illness—to a more proactive model. This new model aspires to go beyond curing diseases by focusing on prevention, quality of life, personalized care, disease prediction, and placing the patient at the center of care.

This paradigm shift aligns with the principles of 5P medicine: preventive, participatory, personalized, predictive, and population-based. In this context, health language technologies (HLTs) and tools like the one proposed in this project play a crucial role by harnessing the potential of health data collection and the large-scale analysis of health information.

Effective RDI (Research, Development, and Innovation) planning that aligns with both organizational goals and environmental context will ensure that technological advancements are both socially relevant and strategically implemented. This alignment will facilitate the integration of innovations into clinical practice through a productive ecosystem. Moreover, identifying emerging research trends that create new market opportunities can drive the growth of companies positioned to meet these demands. Ultimately, this will lead to tangible improvements in patients' quality of life, while fostering economic growth and societal wellbeing.

Proceedings of Recent Advances in Natural Language Processing,pages 767–770
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-088

2 State of the art

The discovery of knowledge from natural language can be conceptualized as a multi-stage process that transforms raw textual data into a structured and semantically rich representation, typically in the form of an ontology. This pipeline begins with the manual or semi-automatic creation of annotated linguistic resources, which requires the definition or selection of annotation schemes tailored to the target domain. From these annotated corpora, machine learning algorithms can be trained to generalize the annotation process across large-scale unstructured texts.

Once annotated, relevant entities and relationships are extracted and represented in a semantic graph. At this stage, post-processing tasks—such as redundancy elimination, entity normalization, and inconsistency detection—are applied to enhance the coherence of the graph. The result is a unified semantic structure where the implicit knowledge within the original text is made explicit and formalized, often in ontology formats such as OWL or RDF.

Natural language processing (NLP) techniques have been widely adopted for this purpose. Early systems such as ISODLE (Weber and Buitelaar, 2006) illustrate the integration of machine learning with semantic annotation. Rule-based systems like OntoLT (Buitelaar and Sintek, 2004) proposed mappings between syntactic structures and ontology classes, enabling concept and relation extraction. In contrast, statistical and probabilistic approaches emerged to reduce manual intervention. Notable examples include LEILA (Suchanek et al., 2006) and Text2Onto (Cimiano and Völker, 2005), which leverage probabilistic models to identify concepts and instances. The KnowItAll system (Etzioni et al., 2004) introduced pointwise mutual information (PMI) to enhance the relevance of extracted instances.

Further advances sought to go beyond mere extraction by inferring higher-level structures. Systems such as OntoGain (Drymonas et al., 2010) and ASIUM (Faure and Poibeau, 2000) utilized clustering techniques to induce concept hierarchies in an unsupervised manner. These approaches marked a shift toward more autonomous ontology learning. More recently, lifelong learning frameworks such as NELL (Never-Ending Language Learning) (Mitchell et al., 2018) have demonstrated the potential of continuously harvesting knowledge from web data, iteratively refining and expanding semantic structures.

Recent literature also emphasizes the importance of deep learning and transformer-based models (e.g., BERT, RoBERTa) in automating entity and relation extraction with improved accuracy (Devlin et al., 2019; Liu et al., 2019). These models provide contextual embeddings that enhance the semantic understanding of textual data, facilitating more robust ontology population tasks.

In summary, the process of extracting knowledge from text requires the integration of various NLP techniques, semantic representation formalisms, storage and reasoning mechanisms, and quality evaluation metrics. The convergence of rule-based, statistical, and deep learning approaches—combined with knowledge graphs and ontologies—continues to advance the field toward more scalable, adaptive, and semantically aware systems.

3 Data analytics and trends

This task focuses on the development of tools aimed at extracting the maximum amount of statistical information related to document profiles over specific time intervals. The objective is to enable the identification of trends and the analysis of topic evolution within the research areas relevant to the project. To achieve this, machine learning techniques—such as time series analysis—will be employed, along with advanced data visualization methods. These visualizations serve a dual purpose: (1) they support the generation of new knowledge, and (2) they illustrate the temporal and statistical evolution of the information associated with the identified ontology.

The ontology lifecycle encompasses key processes such as creation, management, analysis, and reuse. These processes are structured as workflows composed of multiple activities, defined based on established methodologies for ontology model development. To ensure these activities are effectively executed, they must be supported by appropriate mechanisms and tools. In particular, robust visualization techniques are essential, as they provide interactive interfaces that empower users to abstract, conceptualize, understand, represent, and learn from the underlying knowledge.

A critical component of this task is semantic exploration and recommendation. With the existence of a semantic database, it becomes essential to im-

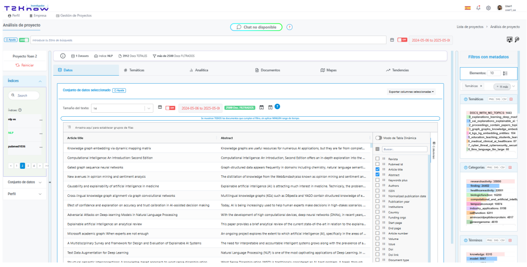


Figure 1: Data collection

plement mechanisms for navigating the semantic network. This will be facilitated through query languages such as SPARQL (https://skos.um.es/TR/rdf-sparql-query/) and Cypher (https://neo4j.com/developer/cypher/). These mechanisms will not only enable document retrieval through metadata filters but also support aggregate queries for trend discovery. Moreover, they will allow for the recommendation of profiles—such as documents, authors, institutions, topics, and named entities—based on the semantic links that interconnect the network.

4 T2KNOW Platform

The T2KNOW Platform is designed to support the exploration and analysis of scientific and technical knowledge across various domains. The figures below illustrate key functionalities of the platform.

Figure 1 shows the data collection process, where structured and unstructured sources are gathered to build a comprehensive dataset. In Figure 2, users can access detailed article data, enabling in-depth examination of individual publications, including metadata and content.

Figure 3 presents the trend analysis capabilities, which allow users to detect emerging topics and monitor their evolution over time. Finally, Figure 4 illustrates the geographical distribution of the articles, providing insights into regional contributions and the global reach of scientific production.

Figure 5 showcases a Retrieval-Augmented Generation (RAG) component integrated into the platform. This feature enables users to interact with the system through natural language queries, leveraging both large language models and the underlying document collection to provide accurate, context-aware answers. By retrieving relevant documents and combining them with generative capabilities, the RAG module enhances information accessibility and supports more intuitive exploration of the knowledge base.

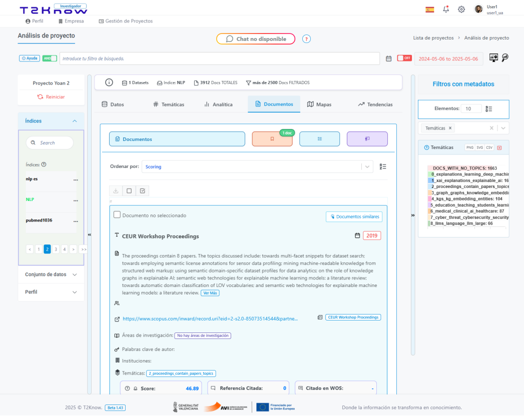


Figure 2: Article data

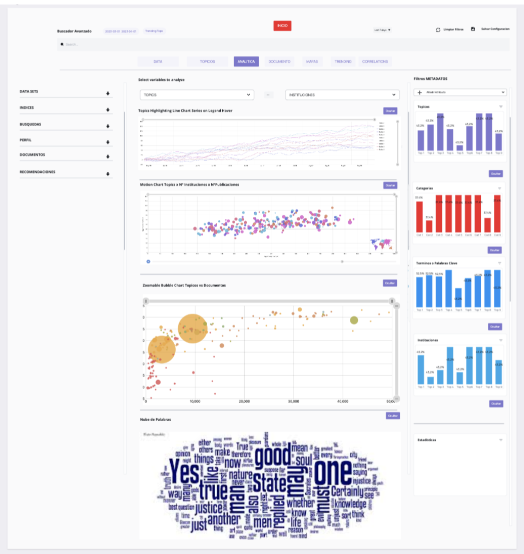


Figure 3: Trend analysis

Together, these components demonstrate the platform's potential to facilitate data-driven research and informed decision-making.

Acknowledgments

This research has been funded by the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport), through the project: NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation (CIPROM/2021/021).

References

Paul Buitelaar and Michael Sintek. 2004. Ontolt version 1.0: Middleware for ontology extraction from text. In *Proc. of the Demo Session at the International Semantic Web Conference*.

Philipp Cimiano and Johanna Völker. 2005. text2onto. In *International Conference on Application of Natu-*

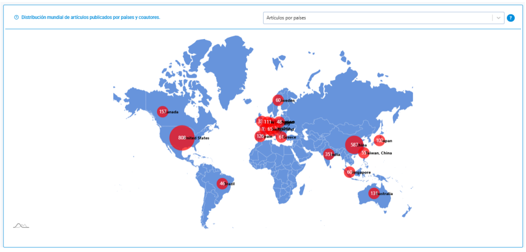
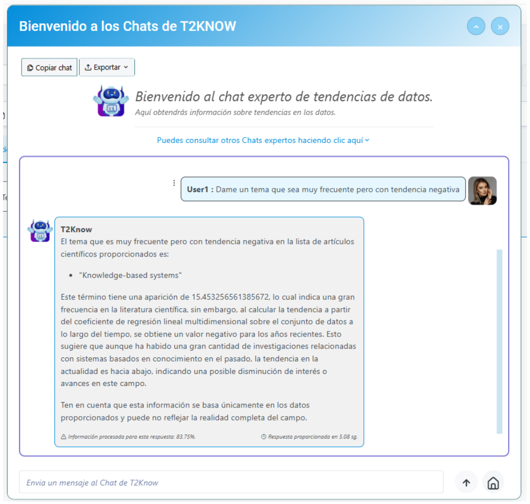


Figure 4: Geographical distribution



Bienvenido al chat experto de tendencias de datos.

Aquí obtendrás información sobre tendencias en los datos.

Puedes consultar otros Chats expertos haciendo clic aquí

User1 : Dame un tema que sea muy frecuente pero con tendencia negativa

T2Know

El tema que es muy frecuente pero con tendencia negativa en la lista de artículos científicos proporcionados es:

- "Knowledge-based systems"

Este término tiene una aparición de 15.453256561385672, lo cual indica una gran frecuencia en la literatura científica, sin embargo, al calcular la tendencia a partir del coeficiente de regresión lineal multidimensional sobre el conjunto de datos a lo largo del tiempo, se obtiene un valor negativo para los años recientes. Esto sugiere que aunque ha habido una gran cantidad de investigaciones relacionadas con sistemas basados en conocimiento en el pasado, la tendencia en la actualidad es hacia abajo, indicando una posible disminución de interés o avances en este campo.

Ten en cuenta que esta información se basa únicamente en los datos proporcionados y puede no reflejar la realidad completa del campo.

⚠ Información procesada para esta respuesta: 83.75%.

Respuesta proporcionada en 5.08 sg.

Figure 5: RAG

ral Language to Information Systems, pages 227–238. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

Euthymios Drymonas, Kalliopi Zervanou, and Euripides GM Petrakis. 2010. Unsupervised ontology acquisition from plain texts: the OntoGain system. In *International Conference on Application of Natural Language to Information Systems*, pages 277–287. Springer.

Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2004. Web-scale information extraction in knowitall:(preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, pages 100–110. ACM.

David Faure and Thierry Poibeau. 2000. First experiments of using semantic knowledge learned by asium for information extraction task using intex. In *Proceedings of the ECAI workshop on Ontology Learning*.

Yinhan Liu et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tom M Mitchell et al. 2018. Never-ending learning. *Communications of the ACM*, 61(5):103–115.

Fabian M Suchanek, Georgiana Ifrim, and Gerhard Weikum. 2006. 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*, pages 18–25.

Nicolas Weber and Paul Buitelaar. 2006. 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.