Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 193-196 recibido 31-03-2019 revisado 06-05-2019 aceptado 15-05-2019

Corpus Viewer: NLP and ML-based Platform for Public Policy Making and Implementation

Corpus Viewer: una plataforma basada en PLN y Aprendizaje Automático para diseño e implementación de política pública

David Pérez-Fernández[1], **Jerónimo Arenas-García**[1,2], **Doaa Samy**[1,3]
Antonio Padilla-Soler[4], **Vanesa Gómez-Verdejo**[2]
[1]Secretaría de Estado de Avance Digital, Spain
[2]Universidad Carlos III de Madrid, Spain
[3]Instituto de Ing. del Conocimiento (IIC), Madrid, Spain and Cairo University, Egypt
[4]Ministerio de Industria Comercio y Turismo, Spain
dperezf@mineco.es, jeronimo.arenas@uc3m.es, doaa.samy@iic.uam.es

Abstract: Corpus Viewer is a production service developed by the State Secretary for Digital Advancement (SEAD) within the framework of the National Language Technologies Plan (Plan TL), promoted by the same State Secretary. Corpus Viewer relies on Natural Language Processing (NLP), Machine Learning (ML) and Machine Translation (MT) to analyze structured metadata and unstructured textual data in large document corpora. The platform allows the decision maker and the policy implementer the possibility of analyze R&D&i information space (mainly patents, scientific publications and public aids) for evidence and knowledge-based policy making and implementation. In this paper, we describe the main functionalities of the platform and enumerate the techniques it is based on, which include a variety of methods like document topic modeling and graph analysis.
Keywords: Topic modeling, Latent Dirichlet Allocation (LDA), graph analysis, Document Similarity, Automatic Classification, dynamic topic analysis

Resumen: Corpus Viewer es un servicio en producción desarrollado por la Secretaría de Estado del Avance Digital dentro del marco del Plan de Impulso de Tecnologías del Lenguaje (Plan TL). Se basa en técnicas de Procesamiento del Lenguaje Natural (PLN) y Aprendizaje Automático para analizar datos estructurados y no-estructurados en grandes colecciones de documentos como las patentes, las publicaciones científicas de acceso abierto, los proyectos europeos, etc. El objetivo es ofrecer al decisor político y al gestor la posibilidad de navegar en el espacio de la información teniendo una visión de conjunto que le ayude a tomar decisiones basadas en conocimiento y evidencias. En este artículo, se describen las funcionalidades básicas de la plataforma enumerando las técnicas empleadas que incluyen, entre otros, modelados de tópicos y análisis de grafos.
Palabras clave: Modelado de Tópicos, Latent Dirichlet Allocation (LDA), Análisis de Grafos, Similitud entre Documentos, Classificación Automática, Modelado Dinámico de Tópicos.

1 Introduction

In this paper, we present Corpus Viewer, a production service developed by the State Secretary for Digital Advancement (SEAD) within the framework of the National Language Technologies Plan, promoted by the same State Secretary. The development of this service started in 2016 and is still progressing based on the collaboration of several subcontracted University research groups and companies. Corpus Viewer on its 1.0 version, is currently used by three public administrations: SEAD (Ministry of Economy), the Spanish Foundation for Science and Technology (FECYT) and the State Secretary for University and Research, Development and Innovation (SEUIDI) at the Spanish Ministry of Science.

Corpus Viewer relies on Natural Language Processing (NLP), Machine Learning (ML),

ISSN 1135-5948. DOI 10.26342/2019-63-28

and Machine Translation(MT) to analyze structured metadata and unstructured text data in large quantities of documents. Although a generic platform that can be exploited with virtually any collection of text documents, the current deployment of the platform mainly hosts R&D related text corpora, such as patents, scientific publications and projects funded at national, European (Cordis) and international level (NSF, NIH). These data sources are processed to assist in the definition and implementation of R&D&i public policies through a set of functionalities allowing to:

1. compare R&D&i funding and knowledge areas in different geographic regions,
2. identify competitive advantages between countries, regions, organizations,
3. identify R&D&i knowledge areas, as well as their emergence, evolution and even hybridization with other knowledge areas (it provides also metadata aggregation and BI type dashboard visualization),
4. R&D agent (organization, researcher and firm) profiling and,
5. assist in the assessment of the impact of public policies by tracking the outputs of grants, short and long term outcomes in terms of lead-lag.

Corpus Viewer also provides tools for policy implementation, in particular for the selection of evaluators or the retrieval of relevant documents (patents, scientific publications, R&D aid grants and proposals, etc) for innovation evaluation. Furthermore, it is used for plagiarism detection, identification of cases of double funding and fraud in aid grants and proposals submitted for national funding.

In this paper, we briefly describe the basics of the Corpus Viewer platform and announce also some of the newest functionalities that have already been completed and will be incorporated to version 1.5 to be deployed this year. In the next section we briefly describe the main functionalities offered by the platform, whereas Section 3 enumerates some of the most relevant techniques that constitute the artificial intelligence core of the platform.

2 Corpus Viewer functionalities

Current version of Corpus Viewer provides the following functionality for public policy design and implementation:

1. Scalable NLP pipeline and Machine Translation of large volumes of documents.
2. Automatic classification of documents according to available taxonomies using deep learning networks.
3. Topic modeling analysis of document collections and topic inference for new documents.
4. Information retrieval system based document similarity. This function is used to provide identify highly similar documents for plagiarism detection or to avoid project reevaluation.
5. Optimized columnar and textual indexing for efficient searches and query metadata filtering.
6. Thematic correlation and main semantic area detection. Semantic grouping, using semantic graph algorithms, by area of knowledge.
7. Tracking of semantically alike documents (semantic clusters) (emergence, evolution and hybridization with other clusters.
8. Dynamic topic analysis and temporal thematic evolution. Temporal analysis by areas of knowledge, lead-lag between different types of document corpus (Figure 1).
9. Topic-enhanced dashboards to analyze R&D&i distribution by metadata, including geographic area (Figure 2).
10. Automatic profiling and disambiguation of R&D&i key players (researchers, institutions, groups) based on their R&D production.
11. Analysis of collaboration networks between key players in the R&D&i production space.

3 Facilitating techniques

In order to provide the previous functions, the platform heavily relies on a number of NLP and ML techniques:

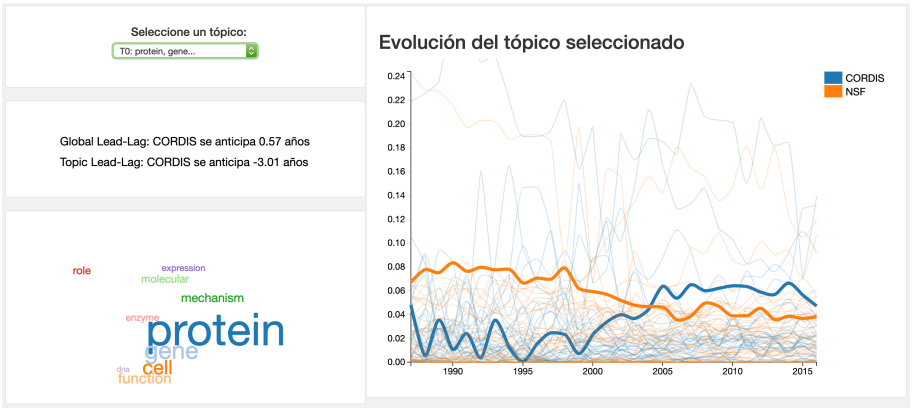

Figure 1: Time evolution and comparison between number of granted projects by topic in FP7/H2020 and NSF.

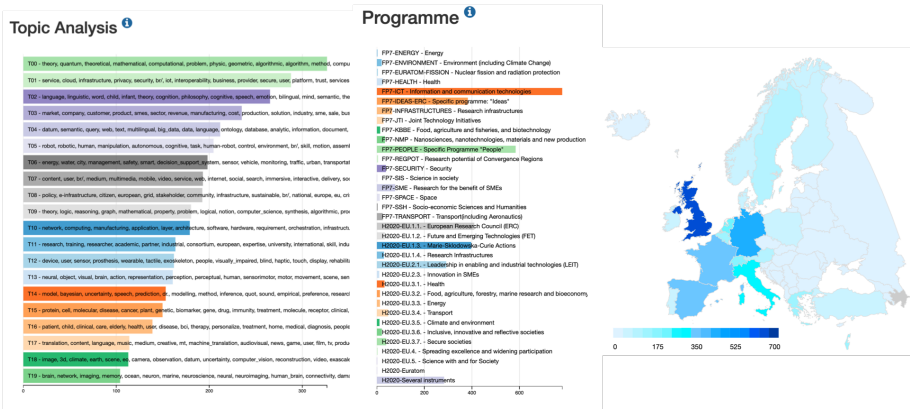

Figure 2: Topic-enhanced dashboard for analysis of the distribution of Cordis projects in the area of Artificial Intelligence.

1. Scalable tokenization, PoS tagging, lemmatization, disambiguation and wikification for English and Spanich languages.
2. Automatic translation (ES-EN), NMT based.
3. Topic modeling, including: 1) static models, such as Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003) and Correlated Topic Models (CTM) (Blei and Lafferty, 2005), 2) dynamic models using the Dynamic Topic Model (DTM) of (Blei and Lafferty, 2006), and a new tool for time analysis developed within the scope of the project and based on the work (Greene and Cross, 2017), and 3) hierarchical LDA and recursive LDA, developed within the scope of the project, to allow multi-level navigation in large corpora.
4. Textual search and document similarity implemented for topic, bag of words, and word Embeddings document representations (Mikolov et al., 2013).
5. Analysis of graphs; modularity, distances between clusters and centrality calculation

Such analysis techniques are based on underlying topics, conceptual indexing of documents, word Embeddings and other techniques of documentary representation. The relations of semantic similarity among the document representations (scientific literature, patents and funded project proposals and grants) equip this space with a metric that favors representation and navigation. Similar documents concentrate in semantic aggregates (document clouds are stable under metric modifications). Semantic aggregates can be obtained using graph analysis tools –see Fig. 3.

Unlike traditional statistical description methods, our approach does not require predefined taxonomies or controlled vocabular-

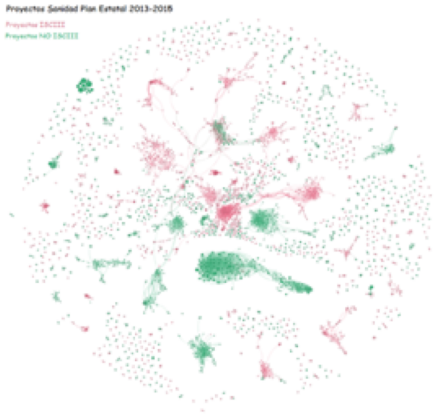

Figure 3: Graph of the Spanish health R&D project proposals and grants based on topic analysis, topic vector interdistances and graph community detection. Different colors represent proposals funded by ISCIII (a health-funding institution, red) and other Spanish R&D funding organizations (green).

ies. Among the advantages of this content-based approach are: increased flexibility for comparing documents from different corpora (without common taxonomies), no need for predefined taxonomies (especially important in wide and dynamic R&D sectors), and the possibility to represent and characterize R&D agents (researchers, research centers, firms) using their R&D production.

4 Use cases examples

In this Section we provide examples of Corpus Viewer use cases and visualizations:
Use-case 1: Comparing the funding destined to certain thematic fields in different countries and its evolution within a certain time frame is of high importance to R&D&i policy makers. Figure 1 illustrates an example of what Corpus Viewer can offer in this use-case. This would provide data-based evidence to answer questions such as:

- How many projects were funded in a certain knowledge area (previously automatically detected) through the framework programmes FP7 and H2020?
- How many projects in the same area were funded by the National Science Foundation within the same time frame?

Use-case 2: Given a certain knowledge field, e.g. Artificial Intelligence, the policy maker would like to know better the investment, and answer questions such as:

- What are the field subareas and how does the invested money distribute among them?
- What countries receive the largest funding in total and in each detected subarea (see Fig. 2).

Use-case 3: Figure 3 represents a semantic graph of R&D health projects funded by Instituto de Salud Carlos III (ISCIII) and by other institutions. Using this approach, Corpus Viewer can help answering the following questions:

- How does the thematic distribution of applications compare to the granted projects?
- Are different institutions (or countries, regions)) funding the same subfields? Are there some fields that receive insufficient funding?

Acknowledgments

This work has been carried out in the framework of the Spanish State Plan for Natural Language Technologies. We would like to acknowledge the different organizations that have contributed to the project under the SEAD-SEUIDI-FECYT agreement for Competitive Intelligence: UPM, IIC, UPF, IXA UPV, Elhuyar.

The work of J. Arenas-García and V. Gómez-Verdejo has been partly funded by MINECO projects TEC2014-52289-R and TEC2017-83838-R.

References

Blei, D. M. and J. D. Lafferty. 2005. Correlated topic models. In *Proc. NIPS*.

Blei, D. M. and J. D. Lafferty. 2006. Dynamic topic models. In *Proc. ICML*.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. In *Proc. NIPS*.

Greene, D. and J. P. Cross. 2017. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, 25:77–94.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*.