Evaluation Task-IberRDI 2020

Title: NLP for the R&D&I Domain in Spanish: Multi-task Evaluation of Document Similarity Metrics

Acronym: IberRDI

Aim

IberRDI aims at taking the initiative in organizing an evaluation task targeting the content-based analysis of text related to the research, development and innovation (RDI) production in Spanish language. The task aims at encouraging Natural Language Processing (NLP) groups to process technical and scientific texts and to tackle the challenges encountered in this type of texts. We focus on document from biomedical domains, but the tools and techniques can be expected to be useful for other domains.

The overall aim of IberRDI is to bring together actors across sectors from Academia, Industry (NLP and RDI policy makers and innovation evaluation agents), Public Administration.

The main goal of the task is to evaluate the quality of document similarity metrics through two different subtasks:

1) Proposing a document representation and use it to compute similarities between documents from a homogeneous collection, i.e. texts from the same corpus, e.g. scientific papers
2) Computing semantic similarities between heterogeneous texts, i.e. documents from different corpora, e.g. scientific papers and patents.

Description

Motivation:

The Spanish National Plan for the Advancement of Language Technologies (PTL, Plan de Tecnologías del Lenguaje) is encouraging public administrations to take advantage of the high degree of digitization to develop smart public services based on the application of language technologies. Within the field of Research, Development and Innovation (RDI), digitization is particularly extensive, and large repositories of open data resources (publications, patent records, project proposals) can be processed to identify the structure and dynamics of the RDI activity in Spain.

A key component of smart systems for the analysis of the RDI production is the identification of different type of relations between documents. Beyond metadata available in the datasets (e.g., funding entities for projects, journal categories for papers, etc.), in this task we focus on the automatic extraction of semantic similarity among RDI items based just on the text available in documents. In applications of technological surveillance, efficient similarity metrics can be used to elaborate large semantic graphs linking and grouping the research production from a given topic, a given country or a given organization. Semantic graphs, in conjunction with document metadata as keywords or citation-based graphs, are powerful sources for the

identification of the structure and dynamics of the research activity. Universities, research-funding institutions and organizations, and policy-makers can take advantage of these technologies.

To this aim, we propose a task involving the analysis of three corpora from the Health sector, which is one of the prioritized areas of the PTL (Villegas, 2017). The goal is to explore different metrics to identify similarities between documents that could work efficiently in different circumstances. We propose several tasks related to the direct evaluation of similarity measurements, both for documents taken from homogeneous and heterogeneous collections.

We expect contributions form researchs from different fields, from Natural Language Processing to Machine Learning. The task will be an opportunity to test technologies ranging from text analysis, topic models (Alexander, 2015), and word or doc embeddings (Kusner, 2015) in an application and sector of great interest for the PTL in particular and for the NLP community in general.

Relevance and Novelty

The computation of semantic similarity measurements between words, sentences, paragraphs or documents is a key component for most Natural Language Processing tasks, including text summarization, information retrieval, plagiarism detection or document classification. It is also a major component for the analysis of large document collections. Tasks related to the evaluation of text similarities have been proposed in relation to text retrieval (Aslam, 2013) and plagiarism detection (Potthast, 2013) (Kasprzak, 2009).

Evaluation tasks related to semantic similarity and text classification in Spanish are not new (see the compilation in (Rosso, 2018)). The design of efficient semantic similarity metrics has been the main purpose of many text analysis tasks. The SemEval workshop series has proposed several tasks on Semantic Textual Similarity between sentences (2012, 2013, 2014, 2015, 2016) and tweets (2015), both in English, Spanish or in cross-lingual pairs (Agirre, 2016). Similarity values are expected to be maximum in case of meaning equivalence.

Our focus here is in document similarity, and our goal is not to identify meaning equivalence (as in a plagiarism detection competitions) but thematic similarity. The present task focuses also on RDI texts at the different stages of the RDI process: project proposals, scientific papers, and patent applications. To the best of our knowledge, there are no previous evaluation tasks centered on the semantic similarity among RDI texts in Spanish language.

Evaluation measures & Methodological Aspects

Although the use of a gold standard based on a the human judgment of similarities between pair of documents presented to several annotators is a common approach in other text similarity tasks (see, for instance, SemEval2012_task_6), manual annotation is an expensive process, because the amount of required labels may grow quadratically with the size of the target corpus and, also, because labelling may require redundancy of annotator to correct annotation biases and detect and correct annotation discrepancies. These scalability and subjectivity issues are especially critical in the context of RDI texts, since semantic similarity

annotation requires the participation of experts in the different fields, and each expert criteria depends on his/her personal research background.

For the previous reasons, in this task we follow an alternative approach based on using information already available in the different collections in the form of citations, keywords, and/or category classification. Apart from being already available, these sources of information have the advantage of being generated directly from authors and experts with a knowledge of the document and the state of the art, which may be a difficult task to a non-expert annotator. Other advantages are quantity and coverage (i.e., they are available for most of the collection items).

For each of the proposed subtasks, a gold standard will be generated using metadata available from the different corpora in the form of a sparse document similarity graph. The goal of participating teams will be to propose a document representation and a metric to compare any two documents in this representation. Such metric will be used to predict the normalized weight (in the range [0, 1]) of the links between each pair of elements using just the textual description of the items. The reference graphs will be partitioned in two subgraphs with disjoint nodes, named the train and test graphs. Only links from the train subgraph will be provided to the participating teams. Evaluation will be based on the comparison of the similarity graphs provided by participants with the reference graph for the test partition.

In summary, for each subtask the following information will be provided to participants:

- Text description for all documents, both in the train and test partitions
- Computed similarities for the documents in the train partition using available metadata as described below. As previously described, the provided similarity graph will be sparse, i.e., there will be pairs of related items for which a similarity value is not provided.
- Evaluation criterion. The quality of a similarity matrix will be computed the cosine similarity function given by

$$L(R,\ G) = \frac{\sum_{i,j} R_{ij} G_{ij}}{\sqrt{\sum_{i,j} R_{ij}^2 \ \sum_{i,j} G_{ij}^2}}$$

where R is the reference graph and G is the similarity graph computed by the participants

In response to this call, participants will have to submit:

- A description of the NLP pipelines.
- The intermediate representation of documents after the NLP pipelines
- A description of the similarity function based on the intermediate representation of the items
- The estimated semantic similarity between each pair of documents in the test partition.

Target Community & Industrial Take up

The target community is the set Universities, research and innovation funding institutions and policy-makers. Specialists on innovation surveillance are potential users of efficient methods to compute similarities between document in the RDI field.

Related Evaluation Activity

We are not aware of previous evaluation tasks dedicated to the similarity measures between documents, on RDI field in Spanish Language using citation, cocitation and metadata similarity goldstandar.

Use cases

There are two main use cases:

Evaluators of innovative projects must assess the degree of innovation of an R&D project based on the current state of the art. This task requires the analysis of other similar projects (submitted to the same or other funding bodies), analyzing related patents that could invalidate the business model and, finally, it is interesting to look for related scientific publications, both to analyze the state of the art and/or to select possible project evaluators.

On the other hand, the direction of public policies on innovation areas needs to analyse the strengths and weaknesses of specific sectors (e.g. AI, IoT, blockchain ...), compare these sectors between countries, analyse their temporal dynamics, quantify the relationship between sectors, etc.

The classifications associated with these corpuses (e.g. IPC/CPC patents, MESH/DECS medical scientific publications) do not have the appropriate granularity, they only classify one corpus, they do not show the degree of belonging of documents to the classes, they are rarely annotated with multiple classes, classifications are not often updated (which is a problem in the R&D sector).

For this reason, a fine-grained analysis of the textual content of a heterogeneous collection of innovation document corpus (public aid for innovation, patents, scientific publications) is necessary. The basis of all described use cases is the similarity between documents.
These corpus, in addition to metadata on the authors, dates, classifications, etc., have a rich collection of links or citations between documents.
This task aims to exploit the relationships between peer documents (established by the authors of the documents themselves or by their evaluators) to find the optimal representation of the documents as well as a measure of their similarity.

Public institutions and private organizations could take advantage of efficient methods to compute similarity graphs for evaluation processes.

Previous Editions

Not available

Linguistic Resources

Data gathering (Sources) & Harvesting Procedure

The dataset is gathered from open public data sources on Health Sciences innovation area. Basically, it includes three types of documents covering the whole RDI process:

- Innovative granted projects corpus: a collection of projects from Health Sciences taken from ISCIII (spanish equivalent to US NIH organism) https://portalfis.isciii.es/es/Paginas/Busqueda.aspx, funded by FIS (Fondo de Investigación en Salud). Currently we have a dump of 2607 research projects including the following information that will be used in the campaign: Project Abstract and Keywords (both in Spanish). Patent and scientific publication citations will be searched using NLP methods.

- Scientific publication corpus: documents taken from Scielo, a collection of Ibero-american journals about Health Sciences (Neves, 2016). Currently SCIELO holds 800k+ documents of which 340k+ are available in Spanish language. We will use a dump of XML files, kindly provided by Biblioteca del Instituto de Salud Carlos III (ISCIII) that includes the following information of interest for this campaign: Paper Abstract (multilanguage), Keywords (multilabel controlled vocabulary on Health sciences, MESH/DECS), Scielo thematic area (9 categories), country of publication, year, author, etc.

  For the extraction of the citation information we will further use the Semantic Scholar information (https://www.semanticscholar.org/) that currently indexes around 180M articles including 150k publications from the SCIELO collection, and already provides disambiguated citations from inside the corpus. Semantic Scholar provides a full dump of the corpus freely available for research activities.

- Granted patent corpus: a subset of patent proposals taken from the Spanish patent office (OEMP, Oficina Española de Patentes y Marcas) web service INVENES INTERPAT y LATIPAT (http://consultas2.oepm.es/InvenesWeb/faces/busquedaInternet.jsp). For the task, we already have a dump of thousands of patent grants including Abstract and full text in Spanish, as well as Cooperative and International Patent Classification codes (CPC and IPC, respectively). Also links to scientific publications and other patents is provided as part of patent evaluation report.

Intercopora citations will be completed using existing citations and using NLP title and author disambiguation.

Annotation Procedure

The gold standard will be based on metadata from the available databases. As previously explained, each subtask of the campaign will be associated to a different similarity graph:

1. Semantic Similarity among Homogeneous documents:
    a. Innovative granted projects corpus: The reference similarity among each pair of project proposals will be based on the available keywords for each project. Especifically, we will calculate the similarity using an extension of Kessler's similarity (Gipp, 2014) between publications: the similarity will be computed as

the number of common keywords over the square root of the product of the number of keywords in each project.

b. Scientific publication corpus: The reference similarity among each pair of articles will be based on number of common papers cited by the two articles. We will use Kessler's similarity: the similarity will be computed as the number of common out-citations of the two papers over the square root of the product of the number of citations included by each paper.
c. Granted patent corpus: The reference similarity among each pair of patentes will be based on available IPC classification codes, taking into account the multilabel structure and hierarchical nature of the IPC system (alternative CPC code quality for document similarity will be examined). To be more specific, the similarity between each pair of patents will be computed as the average of the number of hops in the tree for all pairs of IPC codes assigned to each document.

2. Semantic Similarity among Heterogenous documents:
    a. Innovative granted projects vs scientific publications: Since keywords are available for both corpora, we will use the same reference metric as in subtask 1.a.
    b. Scientific publications vs granted patent corpus: Citation information is available for the patent data, including also citations to scientific papers. Therefore, it will be possible to calculate the reference similarities among items from both corpora using the same out-citation scheme proposed for subtask 1.b.

IPR Issues

Open public data.

Training/Validation/Test Sizes

The validation datasets will be collections of pairs of compared document identifiers with their distance values.

The three reference metrics based on direct citation, cocitation and classification distance based on document metadata will be provided. In some corpora these distances are not calculable because no metadata or links between documents are available.

These data will be provided for both homogeneous and heterogeneous corpus.

Tentative Schedule

The key dates of the tentative schedule are the following:

- Call for participation — Feb. 1st, 2020
- Dataset release: — Feb., 10th, 2020
- Release of reference graphs: — Feb., 25th, 2020
- Result submission due — May, 1st, 2020
- Publication of results — May, 15th, 2020

- System-description paper submission: June, 20th, 2020

Organization Committee

- David Pérez Fernández - Coordinator of the Spanish Language Technologies Plan (Plan TL), Secretariat of State for Digital Advancement (SEAD), Ministry of Economy, Spain
- Jesús Cid Sueiro - Universidad Carlos III de Madrid, Spain
- Doaa Samy - Spanish Language Technologies Plan (Plan TL) & Instituto de Ingeniería del Conocimiento, Spain
- Jerónimo Arenas García - Universidad Carlos III de Madrid, Spain
- Joseba Sanmartín Sola - Fundación Española para la Ciencia y la Tecnología, Spain

Experience of the organizing team

The members of the team represent the different sectors: Academia (Universidad Carlos III Madrid), Public Administration (SEAD and FECyT) and Industry (Instituto de Ingeniería del Conocimiento). The team has been working over the last four years in different projects in the sector of Competitive Intelligence which is the area related to Management and Policy Making in RDI. As a result of this four year experience, an open online tool "Corpus Viewer" has been published (Perez-Fernandez et al., 2019). The tool has been developed and used in real case scenarios by FECYT and SEAD among other actors for Policy Making in RDI and fraud detection. The team has carried out two training workshops for users from Public Administration. Based on the feedback received and the needs detected, further developments are taking place and it is one of the reasons behind the idea of proposing this task given that the tool already implemented similarity measures among documents inter-corpus (among heterogeneous documents) and intra-corpus (among documents from the same datasets).
Members representing FECYT are direct users since their institutional mandate focuses implementing RDI policies. Their participation in organizing the task will provide insights on real case scenarios and will contribute in the validation for the gold standard as well as the results.

Contact Person

- Jesús Cid Sueiro - Universidad Carlos III de Madrid, Spain, jesus.cid@uc3m.es

Funding

Plan TL will provide resources organizational logistics for the task

Other Relevant Issues

PTL might invite a speaker(s) to the workshop.

PTL might also provide travel support for the participants of two best-ranked systems.

References

(Agirre, 2016) Agirre, Eneko, et al. "Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation." *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 2016.

(Alexander, 2015) Alexander, E., & Gleicher, M. (2015). Task-driven comparison of topic models. *IEEE transactions on visualization and computer graphics*, *22*(1), 320-329.

(Aslam, 2013) Aslam, J. A., & Frost, M. (2003, July). An information-theoretic measure for document similarity. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 449-450). ACM.

(Gipp, 2014) Gipp, B. (2014). Citation-based Document Similarity. In *Citation-based Plagiarism Detection* (pp. 43-55). Springer Vieweg, Wiesbaden.

(Kasprzak, 2009) Kasprzak, J., Brandejs, M., & Kripac, M. (2009, September). Finding plagiarism by evaluating document similarities. In *Proc. SEPLN* (Vol. 9, No. 4, pp. 24-28).

(Kusner, 2015) Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). From word embeddings to document distances. In *International conference on machine learning* (pp. 957-966).

(Neves, 2016) Neves, M., Yepes, A. J., & Névéol, A. (2016, May). The scielo corpus: a parallel corpus of scientific publications for biomedicine. In *Procs. of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 2942-2948).

(Pérez-Fernández et al., 2019) Pérez-Fernández, D., Arenas-García, J., Samy, D., Padilla-Soler, A., & Gómez-Verdejo, V. (2019). Corpus Viewer: NLP and ML-based Platform for Public Policy Making and Implementation. Procesamiento Del Lenguaje Natural, 63, 193-196.

(Potthast, 2013) Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., ... & Stein, B. (2013). Overview of the 5th international competition on plagiarism detection. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation* (pp. 301-331). CELCT.

(Rosso, 2018) Rosso, P., Rangel, F., Casacuberta, F., Martínez, C.D. et al., Catálogo Tareas de Evaluación, *Plan de impulso de las Tecnologías del Lenguaje*, Universitat Politècnica de València, Dec. 2018

(Rubner, 2000) Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, *40*(2), 99-121.

(Villegas, 2017) Villegas, M., de la Peña, S., Intxaurrondo, A., Santamaria, J., & Krallinger, M. (2017). Esfuerzos para fomentar la minería de textos en biomedicina más allá del inglés: el plan estratégico nacional español para las tecnologías del lenguaje. *Procesamiento del Lenguaje Natural*, (59), 141-144.