Proceedings of the Iberian Health and Food Language Technologies workshop

Held in conjunction with the XXXVII International Conference of the Spanish Society for Natural Language Processing (SEPLN2021)

Martin Krallinger [1], Jocelyn Dunstan [2] and Francisco M. Couto [3]

[1] *Text Mining unit at the Barcelona Supercomputing Center (BSC), Spain*

[2] *Center for Mathematical Modeling, Faculty of Physical and Mathematical Sciences, University of Chile*

[3] *LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal*

September 2021

Overview

One of the most active application domains of language technologies and NLP is health and biomedicine. This scenario has been accentuated even more due to the current COVID-19 pandemic, where more efficient access, processing and exploitation of multiple types of unstructured data sources, including the biomedical literature, electronic healthcare records, clinical trials, patents, news articles or even social media is of critical importance. Not to mention other data sources like clinical practice guidelines, clinical trials documentation, medical questionnaires, medical informed consent documents, etc.). Nevertheless, despite the considerable progress made in the last two years, most of the current resources and research in clinical and biomedical NLP and text mining is still done mostly on data in English. Adapting and applying such tools to healthcare data in Spanish, as well as other languages like Portuguese or Catalan is not straightforward even under the increasing promise of multilingual NLP applications. Among the main driving forces of clinical NLP research and development one can highlight scientific venues in the form of workshops devoted specifically to this topic, such as BioCreative, BioNLP, AMIA NLP workshops, i2b2, (and its successor n2c2). Other scientific venues like LOUHI, HealTAC or eHealth CLEF, although theoretically being multi-lingual, did only marginally cover content in Spanish. Considering the number of native Spanish speakers and the large community of healthcare-related professionals it is key to provide a more stable and specific scientific environment to present and exchange NLP research results, resources and tools not only to NLP professionals but also to the industrial sector, pharma industry, healthcare professionals including medical informatics and healthcare technology experts. The objective of this workshop is to foster research both at the theoretical as well as at the level of practical real-world applications of NLP technologies applied to a diversity of textual data types. This workshop will also specially cover contributions from industry and companies working on healthcare textual data and results obtained by the pharmaceutical industry.

Finally, in addition to the healthcare NLP session, we will also include a second session covering language technologies applied to the domain of food, nutrition and agriculture. Despite the economic importance of this sector for Spain and Latin-American countries, the potential of NLP applications was only marginally explored.

September, 2021

Málaga

Martin Krallinger

Jocelyn Dunstan

Francisco M. Couto

Topics

- Natural language processing of healthcare text including EHRs, literature, patents, clinical trials, clinical practice guidelines, social media, health forums, or informed consents
- Information extraction of clinical variables, named entities, values and clinical modifiers like negations
- Speech analytics and voice processing for healthcare applications
- Medical controlled vocabularies and ontologies
- Machine-learning and transfer learning for healthcare text analytics
- Information fusion, integration and semantic interoperability of NLP results with structured data or image data
- Explainable models and interpretability of clinical NLP results
- Real-world application and use cases of healthcare text analytics
- Evaluation and benchmarking of clinical NLP methods
- Knowledge discovery and hypothesis generation for health, agriculture, food or nutrition applications
- Information fusion, i.e. integrating data from various sources, e.g. structured and narrative documentation
- Anonymization of sensitive health data
- NLP applied to agriculture, food (food poisoning, allergies, food-medication interactions food contamination, etc....), nutrition
- Multilingual health, agriculture or food NLP approaches (methods, applications, annotated data and terminologies).
- Shared tasks and evaluation campaigns
- Machine translation of specialized content and document types applied to health, agriculture and food/nutrition domains.

Organizers

Martin Krallinger is head of the Text Mining unit at the Barcelona Supercomputing Center (BSC), and former head of the Biological Text Mining unit of the Spanish National Cancer Research Centre (CNIO). He is also responsible for the health and biomedicine related activities of the Spanish Plan for the Advancement of Language Technology (Plan TL) at BSC. Mas has been working in the field of biomedical text mining for over 15 years, on the development of text mining applications for drug-safety, molecular systems biology or oncology. he was involved in the implementation and evaluation of biomedical named entity recognition components, information extraction systems and semantic indexing of large datasets of heterogeneous document types (research literature, patents, legacy reports, European public assessment reports). His research interests, besides clinical NLP include text-mining assisted biocuration, interoperability standards and formats for biomedical text annotations (BioC) as well as development of efficient text annotation infrastructures. Moreover, he was involved in the implementation of the first biomedical text annotation meta-server (Biocreative MetaServer - BCMS) and the follow up BeCalm/TIPS metaserver. He is also particularly interested in the creation of Gold Standard datasets and annotation schema and their use for training supervised machine learning approaches and exploitation at community evaluation shared tasks and evaluation campaigns. Martin Krallinger is one of the main organizers of the BioCreative community assessment challenges for the evaluation of biomedical text mining systems and I participated in the organization of other biomedical NLP shared tasks in including IberEval (BARR, BARR2), IberLEF (MEDDOCAN), eHealth CLEF (QA4MRE Alzheimer's challenge) and BioNLP-OST (PharmaCoNER).

Jocelyn Dunstan. Hello! I'm a physicist working in public health, especially interested in using machine learning and natural language processing to solve key problems. My research focuses on clinical text mining, patient prioritization, and understanding the connection between obesity and food sales. I also lead projects with the industry. I am assistant

professor at the Faculty of Physical and Mathematical Sciences, University of Chile, and a researcher in the Center for Mathematical Modeling (CMM). The webpage of my group is pln.cmm.uchile.cl .

Francisco M. Couto is currently an associate professor with habilitation at FCUL. He graduated (2000) and has a master (2001) in Informatics and Computer Engineering from the IST. He completed his doctorate (2006) in Informatics, specialization Bioinformatics, from the Universidade de Lisboa. He was an invited researcher at EBI, AFMB-CNRS, BioAlma during his doctoral studies. His main research contributions cover several key aspects of bioinformatics and knowledge management, namely in proposing and developing: various text mining solutions that explore the semantics encoded in ontologies; semantic similarity measures and tools using biomedical ontologies; and ontology and linked data matching systems. In 2019, he published a book entitled Data and Text Processing for Health and Life Sciences that provides a step-by-step introduction on how shell scripting can help solve many of the data and text processing tasks that Health and Life specialists face everyday with minimal software dependencies. An adaptation of the book in Portuguese entitled Introdução à Bioinformática Via Linha de Comando was also published in 2019. The book is particularly relevant to Health and Life specialists or students that want to easily learn how to process data and text, and which in return may facilitate and inspire them to acquire deeper bioinformatics skills in the future. He received the Young Engineer Innovation Prize 2004 from the Portuguese Engineers Guild, and an honorable mention in 2017 and the prize in 2018 of the ULisboa/Caixa Geral de Depósitos (CGD) Scientific Prizes.

Contents

Session I

Welcoming to the Iberian Health and Food Language Technologies and BSC/Plan TL resources

by Martin Krallinger

Abstract. During the welcome session, the importance, current research and existing resources for NLP and text mining applications applied to health, nutrition, and food sciences will be summarized together with a short outline of the workshop purpose, content and talks. Some of the ongoing and past efforts in promoting the development of health NLP tools, in particular for content in Spanish by the Plan TL will be presented.

Affiliation. Barcelona Supercomputing Center, Spain

Mining the Sociome for Insights on Chronic Conditions and Healthy Lifestyle

by Anália Lourenço

Abstract. The environmental and lifestyle impacts of diet-related and chronic diseases is a main, timely health topic. Notably, research on diabetes, gluten intolerance, Crohn's syndrome or colorectal cancer attracts considerable attention. At the same time, social media have become an important vehicle to disseminate health information as well as to gather unbiased reports on personal and professional health experiences. Understanding how patients and relatives, health workers and the general public perceive the problematic of these conditions and interact through social media can be key to raise awareness and promote healthy lifestyles.

Our work is focused on the application of machine learning and graph analysis methodologies to the mining of the sociome. Apart from ontological resources, we are exploring the use of deep learning techniques for entity recognition, namely the extraction of disease symptoms, foods and diets. In addition, we are interested in conversational contexts rather than post-level analysis. Understanding social interaction through conversations or as a post-comment relationship enables a more comprehensive depiction of user behaviour, namely in terms of user influence and support.

This talk summarises the lessons we have learned in different works and introduces our new research project on the application of machine learning-based models to detect and curb Health misinformation on social media.

Following recent efforts of the Plan TL on processing medical and social media texts in Spanish, we will also take this opportunity to prompt discussion on current limitations and challenges on mining the sociome for Spanish and Galician Health-related contents.

Affiliation. Universidade de Vigo, Spain

Food Nutrition and Security Cloud: A case usage of NLP technologies to extract food–drug interactions from scientific and clinical texts.

by Enrique Carrillo de Santa Pau

Abstract. Food Nutrition and Security cloud (FNS-Cloud) is a European project bringing together 34 organizations from 14 European countries, trying to solve fragmentation, lack critical mass and unequal access to food and nutrition data. Thus, the objective of FNS-Cloud is to develop an infrastructure of services in the cloud, to manage data on food and nutrition, and safety in a more efficient way, which allows health professionals or researchers to consult them quickly.

FNS-Cloud will use and develop text mining applications to organize and distribute food knowledge from structured and unstructured data, curate and annotate with descriptors, matched with other data, classified, and described in different ways. To annotate FNS data with descriptors, pre-processing will be based on state-of-the-art Natural Language Processing and Machine Learning methods. Data structuring will also be applied to extract semantics used for curation and annotation in a form of descriptors, enabling data linkage and harmonisation.

Herein, we will present our work in the FNS-Cloud project developing an NLP pipeline to extract food-drug interactions (FDIs) mentioned in scientific article abstracts, with the intent of creating resources for clinicians and nutritionists to use when searching for FDIs. Food-drug interaction extraction is formulated as a relationship extraction problem. First entities related to food, drugs and food chemicals are recognized in a sentence and for every pair of food-drug entities it is determined if a relationship of "interaction" exists between them. The pipeline's main hurdle is the poor amount of labelled text data for the food domain, which it overcomes with the combination of labelled data from different sources, ontologies, and food composition databases. The pipeline under development uses a mixture of deep learning, dictionary-based and ontology-based approaches for the recognition and linking of food, chemical and drug entities in the text, and in linking chemicals to foods. Recognition of abbreviations and co-reference resolution approaches are also used to link all different expressions of the same entity together. A transfer learning approach is used to recognize FDI, by training a deep learning relationship extractor to recognize anonymized drug-drug interactions on a corpus of drug-drug interactions and using it on anonymized food-drug interactions.

Health and nutrition are intimately related. The consumption of specialized food with the goal of improving health, preventing disease or as alternative complementary treatment has risen in the last years increased the risk of potential FDIs. Therefore, it becomes essential the development of resources with high standards and quality data in a centralized way to provide health care workers with the best information and tools for giving advice about FDIs.

Food Nutrition Security Cloud (FNS-Cloud) has received funding from the European Union's Horizon 2020 Research and Innovation programme (H2020-EU.3.2.2.3. – A sustainable and competitive agri-food industry) under Grant Agreement No. 863059

Affiliation. IMDEA Food Institute, Madrid. Spain

Use of NLP and Text Mining for health, nutrition, and food: Plan TL/BSC resources, components, corpora and use cases

by Antonio Miranda/Eulalia Farre

Abstract. There is a pressing need to generate more efficient access to food and nutrition-related information applied to health-related content through text mining and natural language processing technologies, not only for data in English but also for other languages like Spanish. For instance, the recent COVID-19 pandemic has also caused noticeable changes in food consumption patterns with potential effects on population health and wellbeing. Most of the previous food-related NLP applications were applied to gastronomy, processing menus, ingredients, and recipes, with far less research on actual clinical and medical application scenarios and content types. The BSC Text Mining Unit, in the context of the Spanish National Plan for Advancement of Language Technologies (Plan TL), has characterized a set of key concepts and entity types of relevance for health and clinical food language technology applications such as food safety (food poisoning, contamination, allergies, food-intolerance, food-drug interactions, food-borne diseases, and patient dietary records). This talk will summarize resources, annotated data types, corpora, and components generated for medical data in Spanish, with potential adaptation to other languages and other application fields (veterinary medicine, agriculture, and environmental health). Particular emphasis will be placed on a novel annotated dataset for the extraction, recognition, and normalization of species mentions, one of the critical concept types for clinical food sciences, and its use for a community evaluation shared task. We will also present the Spanish Food and Health corpus, a dataset annotated with species, diseases, procedures, among other entity types to foster the development of novel language technology applications.

Affiliation. Barcelona Supercomputing Center, Spain

Session II

Food Information Extraction and Normalization: the Past, the Present, and the Future

by Tome Eftimov

Abstract. In the last decades, a great amount of work has been done in predictive modelling of issues related to human and environment health. This is made possible by the existence of several available biomedical vocabularies and standards, which play a crucial role for understanding health information, together with a large amount of health data. In 2019, Lancet Planetary Health noted that the focus of future improvements in our wellbeing and societies will depend on investigating the links between food systems, human health, and the environment. However, despite the large number of available resources and work done in the health and environmental domains, there is a lack of resources that can be utilized in the food and nutrition domain, as well as their interconnections. This talk will provide a summary of the past, the present and the future related to food information extraction and normalization. It will cover different NLP approaches already developed for tracing the food information in textual data, resources that allow making food and nutrition data interoperable, and NLP/ML pipelines for exploring relations between food and biomedical entities. In particular, this is important during the current pandemics of COVID-19, when food provision and security, as well as healthy nutrition and environment, are tremendously needed for quick recovery and long-term sustainable development of our societies.

Affiliation. Gjorgjina Cenikj - Jožef Stefan Institute, Ljubljana, Slovenia

Text mining as a way to select microbial strains to ferment new food products

by Claire Nédellec

Abstract. The online Omnicrobe knowledge base gathers information on food microbiology from multiple sources expressed in natural language, e.g. papers, genetic databases, biological resource center catalogs. The textual data is processed by information extraction methods that automatically identify entities, relationships and categories from relevant taxonomies and ontologies. Food innovation is an example of application. An example of plant juice fermentation illustrates how data linking of worldwide scattered information by NLP and semantic web methods is a powerful solution to speed up food and beverage innovation. Omnicrobe is used as a tool to select a limited set of microbe strains that present all expected properties and availability in Biological Resource Center catalogs. These strains are then tested for their functionalities in different media.

Affiliation. University Paris-Saclay, INRAE

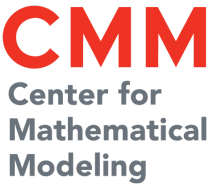
EFSA's initiative for exchanging content and news on Data Science for Food Safety Risk Assessment

by Carsten Behring

Abstract. Combined food safety and data science knowledge is rare across industry, academics and governmental organizations. We need to work closely together to be successful in our aim of making food safety risk assessment more efficient.

Affiliation. European Food Safety Authority

IberHeLT 2021 Sponsors



We acknowledge the Encargo of Plan TL (SEDIA) to BSC for promoting and sponsoring of the Iberian Health and Food Language Technologies workshop, part of SEPLN 2021 (XXXVII INTERNATIONAL CONFERENCE OF THE SPANISH SOCIETY FOR NATURAL LANGUAGE PROCESSING)