First Multilingual clinical NLP workshop

Martin KRALLINGER[a], Carlos PARRA CALDERON[b,c], Antonio MIRANDA[a],
Vasiliki FOUFI[d], Mina BJELOGRLIC[e], Ronald CORNET[f] and Fabio RINALDI[g]
[a] *Barcelona Supercomputing Center (BSC-CNS), Spain*
[b] *Hospital Universitario Virgen del Rocío, Spain*
[c] *Servicio Andaluz de Salud, Consejería de Salud, Spain*
[d] *University of Geneva, Switzerland*
[e] *Hôpitaux Universitaires de Genève, Switzerland*
[f] *University of Amsterdam, The Netherlands*
[g] *University of Zurich, Switzerland*

1. Topic

There is an increasing interest in exploiting the content of unstructured clinical narratives by means of language technologies and text mining approaches. Structured clinical information, in the form of codified clinical records relying on controlled vocabularies such as ICD10 is a key resource for statistical analysis techniques applied to patient data. The results of clinical coding are being used, for instance, as aggregated data to analyze retrospective/prospective aspects of information contained in electronic health records (EHRs).

Clinical natural language processing (NLP) and AI-based document indexing can result in tools for automatic clinical coding by exploiting directly the unstructured content of EHRs. Such tools are playing an increasing role to generate results that do complement health informatics approaches focusing on translational medicine challenges, by providing relevant diagnostic information extracted from clinical narratives. This implies that text mining generated clinical coding results can provide a rich clinical context for patient health information necessary for other data analysis processes like bioinformatics and OMICS data exploration.

Due to the large volume of EHRs and its constant growth, the use of automatic systems to assist experts carrying out clinical coding tasks is becoming increasingly relevant to keep up with the pace of newly generated clinical texts. Automatic clinical coding systems represent also a mechanism to improve the coverage by complementing manual clinical coding activities, potentially improving consistency and reproducibility during the transformation process of unstructured content of EHRs into its corresponding structured representations.

A Gold Standard shareable dataset with high quality manual coding done by professionals with experience in clinical coding together with a proper quality and consistency analysis is key to serve as training and evaluation data for automated clinical coding tools. Ideally, such tools should be able to automatically propose clinical concepts associated to ICD10 codes together with evidence text extracted directly from clinical texts to make the manual coding task much more efficient, consistent, systematic, traceable and interpretable. There have been community efforts to develop such clinical coding systems in English (Pestian et al 2007). In addition, in recent years, shared tasks for clinical coding were proposed for other languages such as French (Goeuriot, et al. 2017), German (Dörendahl, et al. 2019) or Japanese (Aramaki, et al. 2016). However, more limited research was done in Spanish, despite the large volume of clinical content generated in this language not only in Europe but worldwide. Initiative exploring multi-lingual automatic clinical coding, or how to adapt and transfer solutions implemented for one language to another are currently clearly underexplored.

Spanish is spoken by more than 572 million people in the world, either as a native, second or foreign language. There is a pressing need to implement tools able to process automatically the humongous amount of non-English EHRs. The volume and growth rate of non-English clinical texts worldwide justifies the need to promote not only the development of new text mining resources, but also to carry out independent evaluation efforts to assure that the quality and used methods are competitive enough to be of practical value. Shared tasks and community evaluation campaigns have played a critical role in promoting the development of new cutting-edge language technologies, their qualitative and comparative evaluation as well as the adaptation of such solutions across different languages and application domains.

Under the principle of learning how to best solve a particular problem, given a dataset in one language and then transferring the underlying strategies to similar problems or data in another language we propose a shared task that will explore the automatic assignment of ICD-10 codes (CIE10 in Spanish) to health-related documents, the *CodiEsp task*. Participating teams will be provided a Gold Standard dataset of clinical cases in Spanish manually annotated by coding experts, sampled into: a training and a development set to construct and fine-tune their automatic coding systems and a blinded test or evaluation dataset, for which they have to return automatic coding results that will be then assessed against the corresponding manual coding answers.

The workshop proposal includes a shared task overview and results talk, followed by flash talks of the most relevant participating systems. Moreover, the workshop proposal will include ***a 30 minutes panel discussion*** with experts and the workshop audience on the role of shared task to promote clinical NLP, automatic clinical document coding, relevant resources, current challenges and strategies to transfer competitive methods from one language to other.

Some of the proposed topics for the discussion are: *generation of shareable data* collections for clinical NLP and *automatic coding systems*, use of *AI and deep learning* methods applied to *clinical text mining*, exploitation of unstructured content of EHRs for *translational medicine*, explainable IA strategies, *evaluation metrics* and scenarios for automatic clinical coding systems, *multilingual clinical coding* strategies and shared tasks. The intended audience for the workshop and shared task are researchers from medical informatics, clinical/biomedical text mining, bioinformatics, data science, artificial intelligence and machine learning communities, since these profiles showed interest in past tracks from related community challenges (i2b2, n2c2, BioNLP or eHealth CLEF).

Due to the increasing relevance of health and biomedicine as an application domain, we foresee that traditional NLP researches will be attending this session to gain insights on current research trends in particular those related to machine learning techniques and deep learning. Finally, due to the commercial value of the delivered resources and the potential market of automatic coding systems, some SMEs or companies specialized in language technologies and AI might participate as well.

The scope and methodologies of CodiEsp and the proposed discussion align well with many of the topic areas proposed for MIE 2020. For instance, within the Biomedical Data, Tools and Methods, natural language processing methods are required and in past related tasks in other languages, novel artificial intelligence and machine learning systems have been proposed. In addition, in the context of Supporting Care Delivery, clinical coding of EHR is a key step for clinical data mining in our path towards personalized medicine.

2. Rationale Outcome

Clinical coding requires the investment of large amounts of time by highly experienced experts. Moreover, it is a necessary step in the organization of the unstructured information stored in EHRs, whose application to predictive medicine has been already shown relevant in several works (Rajkomar, et al. 2018; Shadmi, et al. 2015). In this sense, the Multilingual clinical NLP workshop at MIE2020 will foster the development of clinical NLP and automatic coding systems.

The development and evaluation of such automatic clinical coding systems requires a Gold Standard dataset with the best quality annotations by experienced professionals in an established classification system such as ICD10. This workshop will provide details on the corpus construction and thus promote the development of other resources allowing independent benchmarking of different approaches. Being clinical coding a complex task, it requires the integration different modules and techniques, such as named entity recognition and normalization but also AI-based multi-class hierarchical text classification approaches.

Finally, to build automatic systems that can be integrated in the clinical environment, community should take into account explainability. Explainable (or interpretable) systems allow easier error debugging, improve trust, help in offering better customer service, etc. Therefore, within CodiEsp and the MIE2020 workshop, special attention will be placed to Explainable AI and AI strategies with human in the loop.

3. Programme (EFMI Virtual Meetings): 4 Jul 2020, 13:00 CET

- [13:00 – 13:20] Workshop motivation and CodiEsp task overview.
- [13:20 – 14:00] Flash presentation talks related to clinical NLP and importance of shared tasks and open discussion on the approaches multilingual clinical natural language processing and text mining systems with a selected panel of experts both from academia, public healthcare and industry.