Report of the First International Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages (SIIRH) *

Francisco M. Couto[1][0000−0003−0627−1496] and Martin Krallinger[2][0000−0002−2646−8782]

[1] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
fcouto@di.fc.ul.pt
[2] Life Science Department, Barcelona Supercomputing Centre (BSC-CNS), C/Jordi Girona 29-31, 08034, Barcelona, Spain martin.krallinger@bsc.es

Abstract. This article briefly summarizes the talks and discussions that occurred during the first edition of the International Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages (SIIRH). The workshop was a virtual event held on April 14, 2020 in conjunction with the 42nd European Conference on Information Retrieval (ECIR2020). The article also presents the main conclusions and future perspectives of the field taking into account the discussions that occurred during the event. All the documents and videos related to the workshop are available at the workshop site: https://sites.google.com/view/siirh2020/.

Keywords: Semantic Indexing · Ontologies · Controlled Vocabularies · Information Retrieval · Text Mining · Natural Language Processing · Biomedical Informatics

1 Introduction

The first edition of the International Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages (SIIRH)[2] was a virtual event held on April 14, 2020 in conjunction with the 42nd European Conference on Information Retrieval (ECIR2020). Semantic Indexing and health-related IR are topics of particular interest for the community that participates in the European Conference on Information Retrieval. This is demonstrated by the topics of interest in the call of papers of ECIR 2020 that cover this workshop scope, namely natural language processing and domain specific search.

* Supported by FCT through funding of the DeST: Deep Semantic Tagger project, ref. PTDC/CCI-BIO/28685/2017, and LaSIGE Research Unit, ref. UIDB/00408/2020

Fig. 1. SIIRH2020 Program

Local Time (GMT+1)	Title	Presenter	Timezone
2:00 pm	**Opening Remarks**	Francisco Couto	GMT+1
2:05 pm	**Keynote Talk I** CoronaTracker: A framework for managing and tracking data during crisis.	Cher Han Lau	GMT+11
	MESINESP/PlanTL Session	André Lamurias (session chair)	GMT+1
2:30 pm	**Keynote Talk II** BioASQ: The challenge and the community of biomedical semantic indexing and question answering	George Paliouras	GMT+3
2:55 pm	MESINESP: Medical Semantic Indexing in Spanish: current and future directions	Martin Krallinger	GMT+2
	Open Session I (Full Papers)	Francisco Couto (session chair)	GMT+1
3:15 pm	First Steps Towards Patient-Friendly Presentation of Dutch Radiology Reports	Koen Dercksen	GMT+2
3:25 pm	Enriching Consumer Health Vocabulary Using Enhanced GloVe Word Embedding	Mohammed Ibrahim	GMT-5
3:35 pm	SmokPro: Towards Tobacco Product Identification in Social Media Text	Kartikey Pant	GMT+5.30
3:45 pm	**Keynote Talk III** The COVID-19 Open Research Dataset	Kyle Lo and Lucy Lu Wang	GMT+7
	Open Session II (Short Papers)	Martin Krallinger (session chair)	GMT+2
4:15 pm	Twitter goes to the Doctor: Detecting Medical Tweets using Machine Learning and BERT	Kevin Roitero	GMT+2
4:20 pm	Biomedical Question Answering using Extreme Multi-Label Classification and Ontologies in the Multilingual Panorama	André Neves	GMT+1
4:25 pm	Towards a multilingual corpus for Named Entity Linking evaluation in the clinical domain	Pedro Ruas	GMT+1
4:30 pm	**Closing Remarks**	Martin Krallinger	GMT+2

On March 11, 2020 the organizers decided to transform ECIR2020 as a open live event due to the worldwide COVID-19 situation. Due to this change, SIIRH and all other ECIR workshops were also transformed in virtual events [8]. SIIRH organization decided to have a three hour programme (2pm-5pm) with short presentations in order to maintain the event as most interactive as possible and deal with presenters from different timezones (from GMT-5 to GMT+11) (see Figure 1). Some additional pre-recorded talks with more details about the works were also provided. All videos were made available as a YouTube playlist[3]. The event was held using Zoom and reached more than 50 online participants. The preliminary proceedings were also published online[4].

Given the COVID-19 situation the keynotes talks, and in many other discussions, focused on solutions on how semantic indexing and information retrieval

[3] https://www.youtube.com/playlist?list=PL6RYRv3A1tLwpD4aTbSVraUZNITbkRxZd
[4] https://drive.google.com/open?id=1-sF_0R3uGinq5ybcAM5H54k5yujJE8o6

systems can help in this health crisis. The solutions allow the scientific community to better deal with the huge amount of information that has to be processed and analyzed, to find ways to contain the spread of a virus as soon as possible.

2 Program

The workshop program was divided as following:

- Opening Session
- MESINESP/Plan TL
- Full Papers
- Short Papers
- Closing Remarks

The full program is available online[5].

2.1 Opening Session

This session included initial remarks from Francisco Couto about the organization of the workshop, and the importance of the field, specially in times of world health crisis as we face nowadays.

The session included the keynote talk entitled *CoronaTracker: A framework for managing and tracking data during crisis* by Cher Han Lau. The talk presented the relevance and collaborative work[6] done to track all the information related COVID-19, including a multi-lingual perspective. The video is available on YouTube[7].

2.2 MESINESP/Plan TL Session

This session was chaired by Andre Lamurias and started with a keynote talk entitled *BioASQ: The challenge and the community of biomedical semantic indexing and question answering* by Georgios Paliouras. The talk showed the importance of international challenges, such as BioASQ[8], to improve the performance of current solutions of semantic indexing and question answering. The video is available on YouTube[9].

The session ended with a presentation entitled *MESINESP: Medical Semantic Indexing in Spanish: current and future directions* by Martin Krallinger, where he demonstrated the relevance of the multi-lingual perspective to community challenges on the field. The video is available on YouTube[10].

[5] https://drive.google.com/open?id=1ds5F7WUR5GZAcjKVcEqS-NVuxy14tDLU
[6] https://www.coronatracker.com/
[7] https://youtu.be/DJtbQfke7A0
[8] http://bioasq.org/
[9] https://youtu.be/GtaEtt3OwCY
[10] https://youtu.be/4EsNf_UYheM

2.3 Full Papers Session

This session was chaired by Francisco Couto and included the presentations of the three works that were accepted as full papers.

The first work entitled *First Steps Towards Patient-Friendly Presentation of Dutch Radiology Reports* was presented by Koen Dercksen [3]. The video is available on YouTube[11].

The second work entitled *Enriching Consumer Health Vocabulary Using Enhanced GloVe Word Embedding* was presented by Mohammed Ibrahim [7]. The video is available on YouTube[12].

The third work entitled *SmokPro: Towards Tobacco Product Identification in Social Media Text* was presented by Kartikey Pant [4]. The video is available on YouTube[13].

The session ended with a keynote talk entitled *The COVID-19 Open Research Dataset* by Kyle Lo and Lucy Lu Wang where they showed their large-scale and recent effort[14] in creating a corpus containing information related COVID-19. There was also a discussion about the importance and advantages of including multi-lingual documents on such efforts. The video is available on YouTube[15].

2.4 Short Papers Session

This session was chaired by Martin Krallinger and included the presentations of the three works that were accepted as short papers.

The first work entitled *Twitter goes to the Doctor: Detecting Medical Tweets using Machine Learning and BERT* was presented by Kevin Roitero [5]. The video is available on YouTube[16].

The second work entitled *Biomedical Question Answering using Extreme Multi-Label Classification and Ontologies in the Multilingual Panorama* was presented by André Neves [1]. The video is available on YouTube[17].

The third work entitled *Towards a multilingual corpus for Named Entity Linking evaluation in the clinical domain* was presented by Pedro Ruas [9]. The video is available on YouTube[18].

2.5 Closing Remarks Session

Martin Krallinger ended the workshop summarizing the main ideas discussed during the event, and pointing out future venues for the advancement of the field, and how it can positively impact the health sector.

[11] `https://youtu.be/N461QEG9r3M`
[12] `https://youtu.be/9GfhtivnONQ`
[13] `https://youtu.be/96lFloRSUYI`
[14] `https://pages.semanticscholar.org/coronavirus-research`
[15] `https://youtu.be/geX4hSRW2vA`
[16] `https://youtu.be/I-SzgxU3KdM`
[17] `https://youtu.be/G8YttYTn89Q`
[18] `https://youtu.be/1SE7PY3sFtA`

3 Conclusions

The workshop was a venue for the different types of contributors, mainly task providers and solution providers, to meet together and exchange their experiences. Thus, the main outcome was the gathering of a group of researchers with hands-on expertise on developing Information Retrieval solutions based on semantic indexing for Life and Health Sciences, which together discussed how to define a road map of what challenges the community should address to produce more efficient and robust solutions.

The main challenge in the field is to motivate and encourage more IR researchers to work with heterogeneous health-related content types in multiple languages. Given that, it is critical to provide training corpora and search solutions for non-English content as well as cross-language or multilingual IR solutions [10], as well as exploitation of evaluation settings and data collections generated through these kind of efforts (both during the evaluation period and afterwards) [6]. We expect that further investigation on the topics will continue after the workshop, based on new insights obtained through discussions during the event.

References

1. André Neves, A.L., Couto, F.: Biomedical question answering using extreme multi-label classification and ontologies in the multilingual panorama. In: Proceedings of the International Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages (SIIRH 2020) (2020)
2. Couto, F., Krallinger, M.: Proposal of the first international workshop on semantic indexing and information retrieval for health from heterogeneous content types and languages (SIIRH). In: Proceedings of the 42nd European Conference on Information Retrieval (ECIR 2020) (2020)
3. Dercksen, K., de Vries, A.P.: First steps towards patient-friendly presentation of dutch radiology reports. In: Proceedings of the International Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages (SIIRH 2020) (2020)
4. Himakar Yv, K.P., Mamidi, R.: SmokPro: Towards tobacco product identification in social media text. In: Proceedings of the International Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages (SIIRH 2020) (2020)
5. Kevin Roitero, Cristian Bozzato, V.D.M.S.M., Serra, G.: Twitter goes to the doctor: Detecting medical tweets using machine learning and BERT. In: Proceedings of the International Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages (SIIRH 2020) (2020)
6. Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodrguez, H., Lopez Martin, J., Villegas, M., Krallinger, M.: Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). vol. TBA, p. TBA. CEUR Workshop Proceedings (CEUR-WS. org), Bilbao, Spain (Sep 2019), TBA (2019)

7. Mohammed Ibrahim, Susan Gauch, O.S., Alqahatani, M.: Enriching consumer health vocabulary using enhanced GloVe word embedding. In: Proceedings of the International Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages (SIIRH 2020) (2020)
8. Nunes, S., Little, S., Bhatia, S., Boratto, L., Cabanac, G., Campos, R., Couto, F.M., Faralli, S., Frommholz, I., Jatowt, A., Jorge, A., Marras, M., Mayr, P., Stilo, G.: ECIR 2020 workshops: Assessing the impact of going online. Tech. Rep. arXiv:2005.06748 [cs.IR], arXiv.org (2020)
9. Pedro Ruas, A.L., Couto, F.: Towards a multilingual corpus for named entity linking evaluation in the clinical domain. In: Proceedings of the International Workshop on Semantic Indexing and Information Retrieval for Health from heterogeneous content types and languages (SIIRH 2020) (2020)
10. Villegas, M., Intxaurrondo, A., Gonzalez-Agirre, A., Marimon, M., Krallinger, M.: The MeSpEN resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. In: Proceedings of the LREC 2018 Workshop "MultilingualBIO: Multilingual Biomedical Text Processing", Paris, France. European Language Resources Association (ELRA) (2018)