The Northern European Association for Language Technology

Workshop on NLP and Pseudonymisation Proceedings

Editors: Lars Ahrenberg and Beáta Megyesi

NODALIDA 2019
September 30, 2019
Turku, Finland

Cover image: Turku Castle, Turku, Finland by Joakim Honkasalo @jhonkasalo.
url: https://unsplash.com/photos/OitUG45b51Y

Proceedings of the Workshop on NLP and Pseudonymisation

Editors: Lars Ahrenberg and Beáta Megyesi

September 30, 2019
Turku, Finland

Published by
Linköping University Electronic Press, Sweden
Linköping Electronic Conference Proceedings, No. 166
Series: NEALT Proceedings Series, No. 41

ISSN: 1650-3686
eISSN: 1650-3740
ISBN: 978-91-7929-996-5

Preface

The goal of making research data freely available often comes into conflict with the rights of individuals. These rights are mainly of two kinds: intellectual property rights and rights to personal data protection. In Europe, the rights to personal data protection have been codified in the recently adopted General Data Protection Regulation, GDPR. While research, as a public interest, can process personal data, the GDPR requires appropriate safeguards to be in place. Consent from authors or subjects cannot always be obtained, or be general enough, and in this case pseudonymisation may be applied, with the intended effect that real individuals no longer can be identified from the language data.

Long before the GDPR, personal data protection has been a concern for creators of language corpora, and there exists a body of literature discussing legal and ethical aspects of corpus publishing. When the data is to be changed or masked in some way, the terms used have been anonymisation or de-identification. With textual data, originals are usually kept, however, which means that anyone with access to the originals and their metadata can make the connection with the transformed text and thus with individuals as authors or participants. For this reason we have used the GDPR term and called this workshop 'NLP for Pseudonymisation'.

NLP is affected in two ways by the conflict. First, it uses language data of all kinds to develop systems, and these data may contain sensitive personal data. Second, it may contribute to making the pseudonymisation process more efficient, or even, more safe. We invited submissions on both of these aspects to the workshop.

NLP has been applied to the problem of deidentification of medical texts for quite a long time. Two of the three papers included in these proceedings deal with medical data. Moreover, in medicine, taxonomies of sensitive data categories are well established and annotated data already in existence. Many other fields, however, not least in the Humanities and Social Sciences, are increasingly aiming to share human-generated data and will need to develop tools and processes for this purpose. We hope that future workshops on the theme of NLP and Pseudonymisation will have a wider spread of contributions.

We would like to express our gratitude to the members of the program committee for their valuable advise and review of papers: Hercules Dalianis, Koenraad de Smedt, Cyril Grouin, Dimitrios Kokkinakis, Krister Lindén, Aurélie Névéol, Sumithra Velupillai, Sussi Olsen, Elena Volodina, and Mats Wirén. We gratefully acknowledge financial support for the workshop from Swe-Clarin, the Swedish node of the European CLARIN infrastructure, with long-term support from the Swedish Research Council.

Linköping and Uppsala, August 26, 2019

Lars Ahrenberg and Beáta Megyesi
Program co-chairs

Program Committee

Lars Ahrenberg (program co-chair), Linköping University, Sweden
Beáta Megyesi (program co-chair), Uppsala University, Sweden
Hercules Dalianis, Stockholm University, Sweden
Koenraad de Smedt, University of Bergen, Norway
Cyril Grouin, LIMSI, CNRS, Université Paris-Saclay, France
Dimitrios Kokkinakis, University of Gothenburg, Sweden
Krister Lindén, University of Helsinki, Finland
Aurélie Névéol, LIMSI, CNRS, Université Paris-Saclay, France
Sumithra Velupillai, King's College, London, UK
Sussi Olsen, CST, University of Copenhagen, Denmark
Elena Volodina, University of Gothenburg, Sweden
Mats Wirén, Stockholm University, Sweden

Invited talk

Martin Krallinger

Head of the Text Mining unit, Barcelona Supercomputing Center (BSC), Spain

Abstract

There is an increasing interest in exploiting the content of unstructured clinical narratives by means of language technologies and text mining. To be able to share, re-distribute and make clinical narratives accessible for text mining and NLP research purposes it is key to fulfill legal conditions and address restrictions related data protection and patient privacy legislations. Thus clinical records with protected health information (PHI) cannot be directly shared "as is", due to privacy constraints, making it particularly cumbersome to carry out NLP research in the medical domain. A necessary precondition for accessing clinical records outside of hospitals is their de-identification, i.e., the exhaustive removal (or replacement) of all mentioned PHI phrases.

Providing a proper evaluation scenario of automatic anonymization tools, with well-defined sensitive data types is crucial for approval of data redistribution consents signed by ethical committees of healthcare institutions. Moreover, it is important to highlight that the construction of manually de-identified medical records is currently the main rate and cost-limiting step for secondary use applications.

This talk will summarise the settings, data and results of the first community challenge task specifically devoted to the anonymization of medical documents in Spanish, called the MEDDOCAN (Medical Document Anonymization) task, as part of the upcoming IberLEF evaluation initiative. This track relied on a synthetic corpus of clinical case documents called the MEDDOCAN corpus. In order to carry out the manual annotation of this corpus we have constructed the first public annotation guidelines for PHI in Spanish carefully examining the specifications derived from the EU General Data Protection Regulation (GDPR). From the 51 registered teams, covering participants both from academia and companies, a total of 18 teams have submitted runs for this track. The top scoring runs represent very competitive approaches than can significantly reduce time and costs associated to the access of textual data containing privacy-related sensitive information. This talk will conclude with a summary of the methodologies used by participating teams to automatically identify sensitive information, together with lessons learned and future steps.

Bio

Martin Krallinger is currently the 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 an expert in the field of biomedical and clinical text mining and language technologies and has been working in this and related research topics since more than ten years, which resulted in over 70 publications and several domain specific text mining applications for drug-safety, molecular systems biology and oncology, etc. 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. He also promoted the development of the first biomedical text annotation meta-server (Biocreative metaserver - BCMS) and the follow up BeCalm/TIPS metaserver. He is one of the main organizers of BioCreative community assessment challenges for the evaluation of biomedical NLP systems and has been involved in the organization of text mining shared tasks in various international community challenge efforts including IberEval, IberLEF, and CLEF.

Contents