BITS :: Call for Abstracts 2019 - Poster

Type	Poster
Session	Big Data: Storage, Analysis and Visualization Biological Databases
Title	Sequence Labelling Techniques for Automatically Integrating Unstructured Genomic Metadata
All Authors	Giuseppe Cannizzaro, Anna Bernasconi*, Arif Canakoglu*, Michele Leone*, Mark James Carman

Affiliation

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, Milano, Italy
*these authors contributed equally to this work

Motivation

The recent growth of public repositories of data regarding the human genome has brought new data integration challenges, mostly related to the heterogeneity of databases structures. A great number of experiment metadata, generated by different laboratories and consortia, are mostly in an unstructured free-text form. This lack of agreement has increased the need for the creation of standards for data structures. The process of adapting data and metadata from the available sources to the new standard formats requires automated techniques given the large quantities of existing data.
In the Gene Expression Omnibus (GEO) [1], one of the most important repository for high-throughput gene expression data, metadata associated with genomics samples lacks a formal and unique structure. Since a unique standard is not enforced on submissions, they often contain textual information, which is unstructured or semi-structured. Moreover, in many cases, they employ different values to express the same concept (e.g., "breast cancer", "breast tumor" or "malignant neoplasm of breast"). These aspects, both at the schema and the value level, create problems when extracting the information using existing text mining tools. The aim of this work is to develop an advanced Natural Language Processing tool to extract all significant information from the metadata documents available in the Gene Expression Omnibus.

Methods

We split the problem into two tasks: We first train a sequence labelling model (such as those used for Part-of-Speech POS tagging and Named Entity Recognition NER [2]) to determine those parts of the text that contain information corresponding to the various fields in the target schema. We then train a domain-specific Named Entity Linking (NEL [3]) model to disambiguate occurrences of named entities, such as cancer types or cell-line indicators, to standardized terms from specialized biomedical ontologies.
We use Deep Neural Networks for both the sequence labelling and entity linking tasks. In particular, we fine tune and extend pre-trained neural networks such as BERT [4] and ELMO [5] for our particular tasks, since they provide state-of-the-art performance on NER/NEL benchmarks such as CoNLL-2003 [6].
In order to train the models, we make use of a large dataset of text from the GEO repository in which each informative phrase is annotated with its corresponding entity: cell-line, tissue, organism, etc.
To build this training dataset, we make use of 1) simple rules for automatic labelling, 2) manual annotation (for previously unlabelled attributes), and 3) data augmentation techniques:
A. With Automatic Labelling, we aim to exploit existing tools to annotate phrases that are easily identifiable based on simple text patterns; such as “Age: ”. This method should produce large quantities of training data for the Deep Learning approach, but will (on its own) be insufficient for producing a useful training dataset. The reason for this is that only simple patterns can be reliably detected using rule-based techniques and thus many non-conforming text occurrences will remain unlabelled. A sequence labeller trained on the resulting dataset may, therefore, struggle to generalize to other types of patterns.
B. To annotate the more complicated occurrences in the training data, we will thus need to rely on manual curation techniques. We are currently building a graphical interface tool that will allow for the fast labelling of words/phrases in a given text, allowing the annotator to directly link the highlighted terms to concepts from an appropriate ontology.
C. Given that Deep Learning techniques require large amounts of training data, we intend to investigate also the utility of providing samples generated using data augmentation techniques, such as flipping of term order, insertion of random noise (inserting characters between words), etc.

Results

The proposed pipeline is designed having in mind the specific case of the GEO repository, a large reference database for the genomic community. However, we believe that the chosen approach will be useful in related domains for automated attribute extraction from free text metadata describing medical experiments (beyond genomics).

Once tuned and trained, the resulting model will be able to automatically locate the relevant attributes (e.g., "cell line" of biosample, the "age" of the donor, or the “disease” analyzed in the experiment) in a plain text document and link them to specific items of biomedical ontologies (such as CL, EFO or NCIT).
If successful, this approach should greatly reduce the amount of effort required by researchers to integrate data across various biomedical data repositories. Researchers should thus be able to analyze more experimental data, spending less of their time on the tedious tasks of manual data entry, curation and management.

Info

References
1. Clough, E. and Barrett, T. (2016). The Gene Expression Omnibus Database. Methods Mol Biol. 1418, 93–110.
2. Yadav, V. and Bethard, S. (2018). A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In Proceedings of the 27th International Conference on Computational Linguistics 2145–2158.
3. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran J.R., (2013) Evaluating Entity Linking with Wikipedia, Artificial Intelligence, 194, 130-150.
4. Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
5. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. Zettlemoyer, L. (2018). Deep contextualized word representations. NAACL.
6. Erik F., Tjong Kim Sang, Fien De Meulder (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. CoNLL-2003.

Figure

-

Availability	-

Corresponding Author

Name, Surname	Arif, Canakoglu
Email	canakoglu@elet.polimi.it
Submitted on	29.04.2019