Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220614

Analyzing the Information Content of Text-Based Files in Supplementary Materials of Biomedical Literature

Nona NADERI[a,b,1], Anaïs MOTTAZ[a,b], Douglas TEODORO[a,b,c], and Patrick RUCH[a,b]
[a] *University of Applied Sciences and Arts of Western Switzerland (HES-SO)*
[b] *Swiss Institute of Bioinformatics, Switzerland*
[c] *University of Geneva, Switzerland*

Abstract. We present an analysis of supplementary materials of PubMed Central (PMC) articles and show their importance in indexing and searching biomedical literature, in particular for the emerging genomic medicine field. On a subset of articles from PubMed Central, we use text mining methods to extract MeSH terms from abstracts, full texts, and text-based supplementary materials. We find that the recall of MeSH annotations increases by about 5.9 percentage points (+20% on relative percentage) when considering supplementary materials compared to using only abstracts. We further compare the supplementary material annotations with full-text annotations and we find out that the recall of MeSH terms increases by 1.5 percentage point (+3% on relative percentage). Additionally, we analyze genetic variant mentions in abstracts and full-texts and compare them with mentions found in supplementary text-based files. We find that the majority (about 99%) of variants are found in text-based supplementary files. In conclusion, we suggest that supplementary data should receive more attention from the information retrieval community, in particular in life and health sciences.

Keywords. Text mining, Semantic annotation, Supplementary materials.

1. Introduction

PubMed stores a wealth of supplementary materials; however, the analysis of such materials is mainly ignored in the literature. In this study, our aim is to examine text-based supplementary materials in terms of their semantic contents.

2. Methods

For our analysis, we have randomly selected 500 PMC articles that include text-based files such as spreadsheets in their supplementary materials. To estimate the importance of text-based supplementary files (spreadsheets, docx, and PDFs), we propose to evaluate MeSH terms extracted from them against MeSH terms manually assigned to reflect the manuscript content [3]. Since the goal is to assess the information content (not

[1] Corresponding Author, Nona Naderi, Information Science Department, HES-SO/HEG Geneva, Switzerland; E-mail: Nona.naderi@hesge.ch.

to achieve the state-of-the-art in information extraction), we use MetaMap text processing tool [1] to extract MeSH terms from abstracts, full-texts, and supplementary text-based files. We further examine these files to identify genetic variants. We use a set of regular expressions to retrieve the different formats of substitution variants. Our regular expressions capture standard formats, as defined by HGVS [2], non standard formats as found in the literature, as well as common variant database identifiers.

3. Results

The recall of MeSH terms using only abstracts is 22.2% and it increases to 28.1% by adding the annotations of 1,643 supplementary text-based files, that is a 5.9 percentage point increase in recall (+20% on relative percentage), which is statistically significant. We further analyze the annotations found in the full-texts and the recall of MeSH descriptors is 46.1%. By adding the annotations found in the supplementary text-based files, the recall increases to 47.6%, which is about a 1.5 percentage point increase (+3.3% on relative percentage). We have found 157 publications with at least one variant. Abstracts, full texts, and supplementary text files contain 14, 518, and 86,348 variants, respectively. The huge majority of variants are thus retrieved in supplementary materials, in particular in spreadsheets, representing 99% of variants. An expert has manually verified a set of 50 random variants found in the full-text, 50 random variants in the supplementary data, as well as the 14 variants found in the abstracts. The manual check of their validity confirms that we retrieve mainly real variants with our regular expressions: 100% of analyzed variants in abstracts and spreadsheets are correct. In the full-text, 78% are correct.

4. Conclusion

In this work, we have performed the analysis of the information content in text-based files in supplementary materials of PubMed articles using MeSH terms and variants. Our results suggest that supplementary materials should be considered as a source of information in the development of text mining tools for assigning MeSH terms to biomedical articles and indexing genetic variants in particular for personalized health.

Acknowledgement: This work has been supported by the CINECA project (H2020 No 825775).

References

[1] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010;17(3):229-36.

[2] den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS recommendations for the description of sequence variants: 2016 update. 2016;37(6):564-9.

[3] Kans J. Entrez direct: E-utilities on the UNIX command line. National Center for Biotechnology Information (US); 2020.