Text-based phenotyping with Semantic Deep Learning: setting up experiments for women's health

Mercedes Arguello Casteleiro[1,2], Niamh Joyce[2], Nava Maroto[3], Maria Jesus Fernandez Prieto[4], Tim Furmston[1], Diego Maseda Fernandez[5], Phillip Lord[6], Chris Wroe[7], John Keane[1], Ying Cheong[2] and Robert Stevens[1]

[1] *University of Manchester, UK*
[2] *University of Southampton, UK*
[3] *Universidad Politécnica de Madrid, Spain*
[4] *University of Salford, UK*
[5] *Mid Cheshire Hospital Foundation Trust, UK*
[6] *Newcastle University, UK*
[7] *BMJ, UK*

Abstract

Semantic Deep Learning (SemDeep) aims to combine Semantic Web and Deep Learning research. Vector arithmetic formulas can be applied to neural language models from Deep Learning. We set up experiments to investigate if incorporating prior knowledge (what is known about the disease) into the vector arithmetic formulas may bring a better performance. This paper investigates a SemDeep approach for text-based phenotyping of four health issues affecting women worldwide: menopause, endometriosis, miscarriage, and infertility. The candidates for the disease phenotype are n-grams that can be mapped to SNOMED CT.

Keywords

Embeddings, SNOMED CT, phenotype

1. Introduction

Social media and the biomedical literature remain largely as silos that lack syntactic and semantic interoperability across and among them. Semantic interoperability requires common terminologies/ontologies such as SNOMED CT [1]. We investigate text-based phenotyping for four health issues affecting women worldwide: menopause, endometriosis, miscarriage, and infertility. In this study, the candidates for disease phenotypes are n-grams with dense vectors representations (i.e. static embeddings) learnt from a textual corpus of PubMed articles (i.e. biomedical literature dataset) or online forums (i.e. social media dataset). Domain experts, such as clinicians and biomedical terminologists, can assess if the candidates for disease phenotypes (n-grams) are true positives (tp) or false positives (fp). By mapping the true positive n-grams to SNOMED CT concepts, it is feasible to create SNOMED CT value sets defined extensionally [1], i.e. an enumerated list of SNOMED CT concepts. Cross-comparing the SNOMED CT value sets obtained from the biomedical literature and the social media may allow the identification of healthcare gaps, e.g. under-reported symptoms.

To validate the proposal, we conducted three experiments (EXP1 to EXP3) for text-based phenotyping using static embeddings created with a PubMed dataset of 300K PubMed articles (titles and available abstracts). The experiments applied vector arithmetic formulas to obtain a list of top ranked candidates for disease phenotypes. EXP2 and EXP3 belong to Semantic Deep Learning (SemDeep) as they incorporate prior knowledge into the vector arithmetic formulas. The performance of the experiments can be assessed using different metrics, such as the area under the Receiver Operating Characteristics (ROC) curve, precision calculated as tp/(tp+fp) and the median of the rank.

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, Basel, Switzerland
EMAIL: M.Arguello-Casteleiro@soton.ac.uk

2. SemDeep Experiments for Women's Health

Female-specific health issues such as menopause, endometriosis, miscarriage, infertility are well known. However, there are no dedicated ontologies for them in well-known repositories, such as BioPortal [2] or the Open Biological and Biomedical Ontology Library [3]. We looked into the UK SNOMED CT human-readable subset [4] and the USA Value Set Authority Center (VSAC) [5]. Only menopause has two value sets in the VSAC with more than 10 SNOMED CT concepts [5].

Considering the lack of dedicated ontologies for women's health issues, we decided to use terms appearing in clinical evidence summaries, such as British Medical Journal (BMJ) Best Practice documents (BMJ-BP for short) [6], for our SemDeep experiments.

Figure 1 has an overview of the experiments set up. A brief overview is the following:

- *Creation of embeddings*. We employed word2phrase from word2vec [7] to compute n-grams from 300K PubMed articles (PMSB dataset). There are 423,462 n-grams appearing 5 times or more in the PMSB dataset. The vector representations were learnt with word2vec [7].
- *EXP1*. This is the baseline experiment, taking the disease name as input for the cosine formula. The phenotype candidates are the top-40 ranked n-grams (highest cosine value).
- *EXP2 and EXP3*. These experiments investigate the incorporation of prior knowledge, i.e. terms appearing in BMJ-BP and having with vector representations.
- *Performance metrics*. We calculate the ROC curve, precision and the median of the rank.

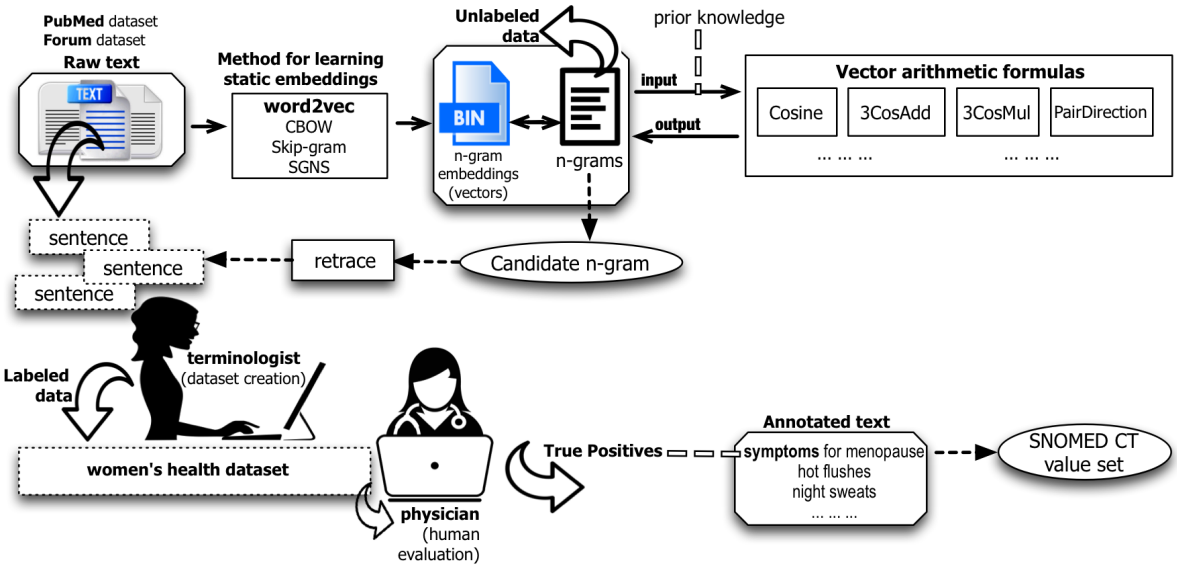

Figure 1: Experiments setup

3. References

[1] SNOMED CT, 2023. URL: http://snomed.org/refsetpg.
[2] BioPortal, 2023. URL: https://bioportal.bioontology.org.
[3] Open Biological and Biomedical Ontology Foundry, 2023. URL: https://obofoundry.org.
[4] UK SNOMED CT subsets, 2023. URL: https://isd.digital.nhs.uk/trud.
[5] USA Value Set Authority Center (VSAC), 2023. URL: https://vsac.nlm.nih.gov.
[6] BMJ Best Practice, 2023. URL: https://bestpractice.bmj.com.
[7] word2vec, 2023. URL: http://code.google.com/p/word2vec.