Deep Learning meets Semantic Web: A feasibility study with the Cardiovascular Disease Ontology and PubMed citations

M. Arguello Casteleiro[1], G. Demetriou[1], W. Read[1], M.J. Fernandez-Prieto[2], D. Maseda-Fernandez[3], G. Nenadic[1,4], J. Klein[5], J. Keane[1,4] and R. Stevens[1]

[1] School of Computer Science, University of Manchester, UK
[2] School of Languages, University of Salford, UK
[3] Midcheshire Hospital Foundation Trust, NHS England, UK
[4] Manchester Institute of Biotechnology, University of Manchester, UK
[5] Institut National de la Santé et de la Recherche Medicale (INSERM) U1048, Toulouse, France

ABSTRACT

Background: Automatic identification of gene and protein names from biomedical publications can help curators and researchers to keep up with the findings published in the scientific literature. As of today, this is a challenging task related to information retrieval, and in the realm of Big Data Analytics.

Objectives: To investigate the feasibility of using word embeddings (i.e. distributed word representations) from Deep Learning algorithms together with terms from the Cardiovascular Disease Ontology (CVDO) as a step to identifying omics information encoded in the biomedical literature.

Methods: Word embeddings were generated using the neural language models CBOW and Skip-gram with an input of more than 14 million PubMed citations (titles and abstracts) corresponding to articles published between 2000 and 2016. Then the abstracts of selected papers from the sysVASC systematic review were manually annotated with gene/protein names. We set up two experiments that used the word embeddings to produce term variants for gene/protein names: the first experiment used the terms manually annotated from the papers; the second experiment enriched/expanded the annotated terms using terms from the human-readable labels of key classes (gene/proteins) from the CVDO ontology. CVDO is formalised in the W3C Web Ontology Language (OWL) and contains 172,121 UniProt Knowledgebase protein classes related to human and 86,792 UniProtKB protein classes related to mouse. The hypothesis is that by enriching the original annotated terms, a better context is provided, and therefore, it is easier to obtain suitable (full and/or partial) term variants for gene/protein names from word embeddings.

Results: From the papers manually annotated, a list of 107 terms (gene/protein names) was acquired. As part of the word embeddings generated from CBOW and Skip-gram, a lexicon with more than 9 million terms was created. Using the cosine similarity metric, a list of the 12 top-ranked terms was generated from word embeddings for query terms present in the generated lexicon. Domain experts evaluated a total of 1968 pairs of terms and classified the retrieved terms as: TV (term variant); PTV (partial term variant); and NTV (non term variant, meaning none of the previous two categories). In experiment I, Skip-gram finds the double amount of (full and/or partial) term variants for gene/protein names as compared with CBOW. Using Skip-gram, the weighted Cohen's Kappa inter-annotator agreement for two domain experts was 0.80 for the first experiment and 0.74 for the second experiment. In the first experiment, suitable (full and/or partial) term variants were found for 65 of the 107 terms. In the second experiment, the number increased to 100.

Conclusion: This study demonstrates the benefits of using terms from the CVDO ontology classes to obtain more pertinent term variants for gene/protein names from word embeddings generated from an unannotated corpus with more than 14 million PubMed citations. As the terms variants are induced from the biomedical literature, they can facilitate data tagging and semantic indexing tasks. Overall, our study explores the feasibility of obtaining methods that scale when dealing with big data, and which enable automation of deep semantic analysis and markup of textual information from unannotated biomedical literature.

*** Contact:** robert.stevens@manchester.ac.uk

1 INTRODUCTION

According to the World Health Organisation cardiovascular diseases (CVDs) are the number one cause of death globally [1]. The SysVASC project [2] seeks to provide a comprehensive systems medicine approach to elucidate pathological mechanisms for CVD, which will yield molecular targets for therapeutic intervention. To achieve this aim it is necessary to gather and integrate data from omics (e.g. genomics, transcriptomics, proteomics and metabolomics) experiments.

The CVD ontology (CVDO) is developed as part of the sysVASC project to provide the infrastructure to integrate omics data that encapsulate findings published in the scientific literature. The CVDO ontology has 172,121 UniProtKB protein classes related to human, and 86,792 UniProtKB protein classes related to mouse. Of these, so far a total of only 8,196 UniProtKB protein classes (i.e. reviewed Swiss-Prot; unreviewed TrEMB; along with Isoform sequences) from mouse and human have been identified as of potential interest to the sysVASC project. An important part of the manually curated effort is to tie experimental findings to the biomedical scientific literature. However, even a project like sysVASC cannot afford to have a team of researchers or curators who can survey the literature regularly and deal with the fundamental task of identifying gene and protein names as a preliminary step to identify the omics information encoded in the biomedical text.

PubMed queries were the starting point of a systematic literature review performed for sysVASC to obtain the omics studies that underpins CVDO and the CVD Knowledge Base (CVDKB). PubMed is a database from the U.S. National Library of Medicine

(NLM) with millions of citations from MEDLINE, life science journals, and online books. In June 2016, PubMed contained 26 million citations with an average of 1.5 papers added per minute [3]. Keeping CVDKB up-to-date is a challenge shared with systematic reviews that aim to keep updated with the best evidence reported in the literature. As of today, searching through biomedical literature and appraising information from relevant documents is extremely time consuming [4,5,6]. Furthermore, omics is a demanding area, where the irregularities and ambiguities in gene and protein nomenclature remain a challenge [7,8]. Krauthammera and Nenadic [9] highlight: "*successful term identification is key to getting access to the stored literature information, as it is the terms (and their relationships) that convey knowledge across scientific articles*". The identification of biological entities in the field of systems biology has proven difficult due to term variation and term ambiguity [10], because a concept can be expressed by various realisations (a.k.a. term variants). A large-scale database such as MEDLINE/PubMed contains longer words and phrases (e.g. *"serum amyloid A-1 protein"*) as well as shorter forms like abbreviations or acronyms (e.g. *"SAA"*). Finding all the term variants in text is important for improving the results of information retrieval (IR) systems like the PubMed search engine, which traditionally rely on keyword-based approaches. Therefore, the number of documents retrieved is prone to change when using acronyms instead of and/or in combination with full terms [11,12].

This paper investigates the feasibility of using Deep Learning, an emerging area of artificial neural networks, for identifying gene and protein names of interest for sysVASC in biomedical text. More specifically, we propose to use the two neural language models Skip-gram and CBOW (Continuous Bag-of-Words) of Mikolov *et al*. [13,14] to produce word embeddings, which are distributed word representations typically induced using neural language models. These word embeddings can be traced back to PubMed citations, and can be also linked to the CVDO classes formalised in the CVD Ontology represented in the W3C Web Ontology Language (OWL) [15].

2 APPROACH

In terms of information/knowledge extraction from texts, over the years, the knowledge engineering (KE) [16] paradigm has lost popularity in favour of the machine learning (ML) paradigm. ML algorithms learn input-output relations from examples with the goal of interpreting new inputs; therefore, the performance of ML methods is heavily dependent on the choice of data representation (or features) to which they are applied [17]. Representing words as continuous vectors has a long history where different types of models have been proposed to estimate continuous representation of words and create distributional semantic models (DSMs). DSMs derive representations for words in such a way that words occurring in similar contexts will have similar representations, and therefore, the context needs to be defined. Some examples of context in DSMs include: Latent Semantic Analysis (LSA) [18] which generally uses an entire document as a context (i.e. word-document models), and Hyperspace Analog to Language (HAL) [19] which uses a sliding word window as a context (i.e. sliding window models). More recently, Random Indexing [20] has emerged as a promising alternative to LSA. LSA, HAL, and Random Indexing are spatially motivated DSMs. Examples of probabilistic DSMs are Probabilistic LSA (PLSA) [21] and Latent Dirichlet Allocation (LDA) [22]. While spatial DSMs compare terms using distance metrics in high-dimensional space [23], probabilistic DSMs measure similarity between terms according to the degree to which they share the same topic distributions [23]. Most DSMs have high computational and storage cost associated with building the model or modifying it due to the huge number of dimensions involved when a large corpus is modelled [29]. Although neural models are not new in DSMs, recent advances in artificial neural networks (ANNs) make feasible the derivation of words from corpora of billions of words: hence the growing interest in Deep Learning and the neural language models CBOW and Skip-gram of Mikolov *et al*. [13,14].

In a relatively short time, CBOW and Skip-gram have gained popularity to the point of being used for benchmarking word embeddings [25] or as baseline models for performance comparisons [26]. We propose applying Mikolov *et al*. [13,14] neural language models, which can be trained to produce high-quality word embeddings on English Wikipedia [25], to automatically extract terms (gene and protein nomenclature) from 14,056,761 free-text unannotated MEDLINE/PubMed citations (title and abstract). Our hypothesis is that word embeddings of high quality should generate useful lists of term variants. As of today, the application of Mikolov *et al*. [13,14] CBOW and Skip-gram to the biomedical literature remains largely unexplored with only some pioneering work [27,28].

3 METHODS

3.1 The CVD Ontology and its Knowledge Base

The CVD ontology (CVDO) provides the infrastructure to integrate the omics data from multiple biological resources, such as the UniProt Knowledgebase (UniProtKB) [29], the miRBase [30] from EMBL-EBI, and the Human Metabolome Database (HMDB) [31]. At the core of CVDO is the Ontology for Biomedical Investigations (OBI) [32] along with other reference ontologies produced by the OBO Consortium, such as the Protein Ontology (PRO) [33], the Sequence Ontology (SO) [34], the three Gene Ontology (GO) sub-ontologies [35], Chemical Entities of Biological Interest Ontology (ChEBI) [36], NCBI Taxonomy Ontology [37], the Cell Ontology (CL) [38], the Uber Anatomy Ontology (UBERON) [39], Phenotypic Quality Ontology (PATO) [40], and Relationship Ontology (RO) [41].

In terms of knowledge modelling, CVDO shares the protein/gene representation used in the Proteasix Ontology (PxO) [42].

3.2 PubMed: from XML to RDF datasets

Through the ftp server from the U.S. NLM we downloaded the MEDLINE/PubMed baseline files for 2015 and also the update files up to 8th June 2016. We created a processing pipeline written in Python that allows the conversion of the downloaded PubMed XML files into W3C Resource Description Framework (RDF) [43] datasets. This pipeline can also be reused to process the results of PubMed searches.

We performed a mapping between the PubMed XML elements [44] and terms from the Dublin Core Metadata Initiative (DCMI), which has been taken up globally and has a publicly available RDF Schema [45].

When pre-processing the textual input for Mikolov et al. [13,14] CBOW and Skip-gram, it is common practice systematically to

lower-case the text and to remove all numbers. However, this is unsuitable when dealing with protein/gene names, because critical information will be lost. To further illustrate this: for human, non-human primates, chickens, and domestic species, gene symbols contain three to six alphanumeric characters that are all in uppercase (e.g. OLR1), while for mice and rats the first letter alone is in uppercase (e.g. Olr1). Therefore, we have introduced some ad hoc rules as part of the pre-processing to guarantee that protein/gene names are preserved.

3.3 Deep Learning: word embeddings using word2vec

This study looks at neural language models, i.e. distributed representation of words learnt by artificial neural networks (ANNs). We adopted the new log-linear models that try to minimise computational complexity. The CBOW and Skip-gram model architecture [13,14] is similar to the probabilistic feedforward neural network language model (NNLM). The feedforward NNLM proposed by Bengio et al. [46] consists of input, projection, hidden, and output layers. In the CBOW and Skip-gram model architecture, the non-linear hidden layer is removed and the projection layer is shared for all the words, so all words are projected into the same position (their vectors are averaged) [14]. The Skip-gram model uses the current word to predict surrounding words, while the CBOW model predicts the current word based on the context.

The basic Skip-gram formulation uses the softmax function [14]. The hierarchical softmax is a computationally efficient approximation of the full softmax. If W is the number of words in the lexicon, hierarchical softmax only needs to evaluate about $\log_2(W)$ output nodes to obtain the probability distribution, instead of needing to evaluate W output nodes.

word2vec [47] is the software package used in this study. It was initially released as open software and is faster than its Python counterpart implementation from Gensim [48]. Using word2vec and out of CBOW and Skip-gram with hierarchical softmax we obtain: 1) a *lexicon* (i.e. a list of terms, typically multi-words) in textual format that is constructed from the input data; and 2) the resulting vectors of the neural DSM in binary mode. In distributional semantics a well-known similarity measure is cosine similarity, i.e. the cosine of the angle between two vectors of n dimensions. If the cosine is close to zero, the two vectors are considered dissimilar, while if it is close to one, this indicates a high similarity between the two vectors.

3.4 Integrating CVDO and word embeddings

The terms from the word embeddings lexicon can be traced back to PubMed citations. Among these terms, there are suitable (full and/or partial) term variants for gene/protein names that can also be linked to the CVDO classes in the CVD ontology. To perform the linkage between word embedding terms and CVDO classes, we looked at the Simple Knowledge Organization System (SKOS) [49], which is a W3C standard aimed at leveraging the power of linked data. In SKOS there are three properties to attach lexical labels to conceptual resources [49]: 1) the preferred lexical label (i.e. skos:prefLabel); 2) the alternative lexical label (i.e. skos:altLabel) for synonyms and acronyms; and 3) the hidden lexical label (i.e. skos:hiddenLabel) for including misspelled variants of other lexical variants or a string for text-based indexing. All of these can be considered annotation properties (i.e. owl:Annotation Property), and allow limited linguistic information only. In this study, we propose using skos:hiddenLabel to store plausible term variants derived from word embeddings for gene and protein classes from the CVD ontology.

3.5 Experimental setup

A gold standard was created using 25 papers that meet the inclusion and exclusion criteria of the sysVASC systematic review performed. The original PubMed query was: "*coronary heart disease AND (proteomics OR proteome OR transcriptomics OR transcriptome OR metabolomics OR metabolome OR omics)*". Out of all the paper abstracts, a total of 107 terms were manually annotated as protein/gene names. Each term was mapped to a CVD ontology class to uniquely identify the conceptual entity (gene/protein) to which the annotated term refers. This can be seen as term standardisation process. Table 1 illustrates the mapping performed.

Table 1. Example of terms from PubMed abstract/title (left column) mapped to labels for protein/gene from the CVD ontology (right column).

Term	(UniProt AC) protein name [gene symbol]
alpha-1-antitrypsin α(1)-antitrypsin	(P01009) Alpha-1-antitrypsin [SERPINA1]
annexin 4	(P09525) Annexin A4 [ANXA4]
superoxide dismutase 3	(P08294) Extracellular superoxide dismutase -Cu-Zn- [SOD3]

The few examples from Table 1 show the lack of standardisation of the field, and illustrate some of the alternative terms from the literature that refer to the conceptual entities (gene/protein) of interest.

In this study we conducted two experiments:

- *Experiment I* – we use the annotated terms from selected papers of the sysVASC systematic review alone (as they appear in the paper abstracts/titles) to obtain the list of 12 top-ranked terms (highest cosine value) from the CBOW and Skip-gram word embeddings. These are candidate term variants.
- *Experiment II* – we enriched/expanded the original annotated terms with terms that appear in the CVDO classes and again produced a list of 12 top-ranked candidate term variants from CBOW and Skip-gram word embeddings.

3.6 Human assessment

Domain experts assessed all the lists of 12 top-ranked candidate term variants obtained for experiment I and II using CBOW and Skip-gram.

3.6.1 Evaluation guidelines We established a strict criterion to mark the list of candidate terms produced by the word2vec word embeddings. Following Nenadic *et al*. [50] a candidate term was marked as *term variant* (TV for short) only when the term fell within the following types of term variation: a) orthographic; b) morphological; c) lexical; d) structural; or e) acronyms and abbreviations. Considering the biomedical value of phraseological expressions (e.g. *"ankyrin-B_gene"* or *"CBS_deficiency"*) we marked them as *partial term variant* (PTV for short); however, they had to refer to the same biomedical concept, i.e. protein or gene name.

3.6.2 Inter-annotator agreement Two domain experts following the above-mentioned evaluation guidelines assigned a simple key code for each candidate term variant: TV, PTV, and NTV (non term variant, meaning none of the previous two categories). The inter-annotator agreement is based on the Kappa measure [51], widely used for inter-annotator agreement on classification tasks; Kappa K is defined as $K = (Pr(a) – Pr(e))/(1 – Pr(e))$, where Pr(a) is the relative observed agreement between annotators, and Pr(e) is the chance agreement.

4 RESULTS

Using a VM with 100 GB RAM and 32 CPU(s) at 4.0 GHz, we obtained the word embeddings from the unannotated PubMed corpus of 14,056,761 free-text MEDLINE/PubMed citations (title and abstract) for Skip-gram (much slower than CBOW) after 17 hours of processing. Due to the lack of space, we show here only some of the results obtained.

Both CBOW and Skip-gram used the same input and generate the same lexicon, however, the resulting vectors of the neural DSM in binary mode were different. Hence, the 12 top-ranked terms for an input query term are likely to differ. For experiment I, only 77 of the 107 terms belong to the lexicon generated. For experiment II, as the CVD ontology is used to provide more context for each term, only 3 terms out of the 107 remained without a valid entry in the lexicon. For experiment I, two domain experts (rater A and B) assessed the 924 pairs of terms corresponding to 77 query terms. For experiment II, there was 87 query terms that include terms from the human-readable labels of key classes (gene/proteins) from the CVDO ontology and considering multiple alternatives, and thus, the same two domain experts assessed 1044 pairs of terms.

Table 2. *Experiment* I: number of terms classified as TV (Term Variants); PTV (Partial TV); and NTV (non TV) by rater A.

Model	Term Variant	Partial TV	Non Term Variant
CBOW	77	93	754
Skip-gram	151	194	579

Table 3. *Experiment II using Skip-gram*: number of terms classified as TV (Term Variant); PTV (Partial TV); and NTV (non TV) by rater A and B.

Domain Expert	Term Variant	Partial TV	Non Term Variant
Rater A	194	240	610
Rater B	161	238	645

Table 2 summarises the number of terms classified as TV, PTV, and NTV for rater A in experiment I using CBOW and Skip-gram. It is easy to derive from Table 2 that Skip-gram is better suited for the task of finding suitable (full and/or partial) term variants for gene/protein names. The observed agreement (i.e. the portion of terms classified as TV, PTV, or NTV on which the two domain experts agree) for experiment I with Skip-gram was 0.80 using weighted Cohen's Kappa measure [51]. Tables 3 summarised the number of terms classified as TV, PTV and NTV for rater A and B for experiment II using Skip-gram. The inter-annotator agreement was 0.74 using the weighted Cohen's Kappa measure [51].

To illustrate qualitatively the results obtained; Table 4 (right column) shows the term annotated (TV, PTV, and NTV) by rater B in experiment II using Skip-gram for *ORL1*, which is a gene symbol. From experiment I using also Skip-gram, no suitable term variants were found. It should be noted that some of the candidate terms listed in Table 4 are well-known aliases of the gene symbol, such as *LOX-1*.

Table 4. *Experiment II using Skip-gram*: 12 top-ranked nearest neighbours by cosine similarity marked by rater B (TV, PTV, and NTV) for the query terms "oxidized_low-density_lipoprotein receptor_1" "OLR1".

Term	Cosine similarity
lectin-like_oxidized_low-density_lipoprotein	0.688603 - **TV**
(LOX-1)_is	0.672042 - **PTV**
atherosclerosis_we_investigated	0.669050 - NTV
receptor-1	0.664891 - NTV
lectin-like_oxidized_LDL_receptor-1	0.663988 - **TV**
lOX-1_is	0.660110 - NTV
human_atherosclerotic_lesions	0.657075 - NTV
oxidized_low-density_lipoprotein_(ox-LDL)	0.655515 - NTV
oxidized_low-density_lipoprotein_(oxLDL)	0.654965 - **PTV**
(LOX-1)	0.652099 - **TV**
proatherosclerotic	0.651571 - NTV
receptor-1_(LOX-1)_is	0.649000 - **PTV**

We observed that some of the protein names annotated from sysVASC systematic review papers, like "annexin 4", can not produce a suitable term variant as they do not appear as such in the generated lexicon generated by CBOW and Skip-gram. However, by enriching them with terms from the CVDO ontology, it is feasible to obtain suitable term variants. For example, "annexin 4" can be mapped to the full protein name "Annexin A4", which has UniProt Accession number P09525 and gene symbol *ANXA4*. Indeed, within the level of observed agreement among the two domain experts, we can safely say that in the first experiment, suitable (full and/or partial) term variants were found for 65 of the 107 terms. In the second experiment, the number increased to 100. Hence, only 7 out of the total 107 remain without suitable (full and/or partial) term variants.

We also observed that the median of the rank (i.e. position in the list of 12 top-ranked terms) for a TV agreed by rater A and B is 3 in both experiment I and II using Skip-gram. In other words, within the level of observed agreement a TV is likely to appear in the first three positions of the 12 top-ranked terms.

5 DISCUSSION

CBOW and Skip-gram have become the state-of-the-art for generating word embeddings. From a quantitative point of view, this study shows that using Skip-gram the number of term variants (TV and/or PTV) for proteins/genes is substantially increased in comparison with CBOW. For experiment II, i.e. when the terms annotated from the sysVASC systematic review papers are enriched/expanded with terms from the CVD ontology, the number of suitable (full and/or partial) term variants for gene/protein increases. The explanation seems quite straightforward as the Skip-gram model takes the word window as a context and predicts surrounding words given the current word [14]. With the aid of the CVD ontology, we can get terms that provide a more pertinent context by: a) enriching a gene symbol with parts of the protein

name; or b) including more than one term related to a protein name. Hence the better results, which in our case means more term variants.

Detecting term variants can be useful for a variety of curation and annotation tasks. As for both experiment I and II, the observed agreed TV are likely to appear in the first three positions of the 12 top-ranked terms; this finding can be the basis of a systematic approach to obtain plausible term variants for the 258,913 UniProtKB protein classes from the CVD ontology. As we proposed in section 3.4, plausible term variants from word embeddings can be easily stored in the CVD ontology by means of the annotation property skos:hiddenLabel. Therefore, when querying the CVD ontology using the query language SPARQL 1.1 [52] for a protein that may appear in the biomedical literature, it possible to use both rdf:label and skos:hiddenLabel. However, the terms stored in the skos:hiddenLabel are more likely to give pertinent results because they are derived from the word embeddings obtained from the 14 million PubMed citations (titles and abstracts), i.e. from the biomedical literature itself. Furthermore, by having transformed PubMed citations into RDF datasets, it is feasible to annotate a PubMed citation not only with MeSH headings/descriptors, or keywords from authors, but also with the terms for the lexicon generated by CBOW and Skip-gram. Thus, it is possible to envision more sophisticated SPARQL 1.1 SELECT queries that are able to retrieve the PubMed citations themselves. Furthermore, from a computational point of view the process described here is affordable and sustainable: new PubMed citations can be converted into RDF on a daily basis; Skip-gram can re-generate the lexicon and the vectors in less than a day for 14 million PubMed citations (titles and abstracts); terms from the human-readable labels of key classes (gene/proteins) from the CVDO ontology can be used as query terms to retrieved the top-ranked terms from the word embeddings re-generated, where the three top-ranked terms (plausible term variants) can be stored as literal values of skos:hiddenLabel. Hence, periodical updates are feasible.

Although text mining technology has made great strides in extracting biomedical terminology from unstructured text sources, the task of normalising (grounding) the extracted terms to commonly used identifiers in ontologies or taxonomies is still quite demanding. Identifying equivalent text realisations for the same biomedical concept can be useful for (i) improving the quality of information in curated resources such as UniProt or the Gene Ontology, and (ii) for linking the information in these resources back to the original text sources; this is helpful when a greater context needs to be explored or for keeping up-to-date with the published literature.

Another potential application is in the area of query expansion for Information Retrieval (IR). Although query enhancement using synonyms is commonly deployed by many of today's IR systems, it is often more difficult to deal with cases of orthographic variations or when new acronyms/abbreviations are introduced for new terms. Identifying term variants can be a way of ameliorating the effect of the classical problem of IR returning either too much or too little for a user query.

Lastly, text mining developers, especially those dealing with rule-based systems, can benefit from unsupervised automated techniques such as the one described in this paper, for building terminological resources from large untagged corpora. Such resources include both terminology lexica and grammars, either manually developed or compiled via grammar induction techniques. The usefulness of this approach for specific annotation tasks will be the subject of future work.

From a research perspective, data integration is not a new challenge in the life sciences. Gomez-Cabrero et al. [53] state: *"there is a need for improved (and novel) annotation standards and requirements in data repositories to enable better integration and reuse of publically available data"*. To the best of our knowledge, this is the first time that word embeddings from Deep Learning and an ontology in OWL have been put together with the aim of linking ontology classes to terms derived from a large corpus of biomedical literature in an unsupervised way and without the need of having the corpus annotated.

6 CONCLUSION

This study demonstrates the benefits of using terms from the CVDO ontology classes to obtain more pertinent term variants for gene/protein names from word embeddings generated from an unannotated corpus with more than 14 million PubMed citations. As the terms variants are induced from the biomedical literature, they can facilitate data tagging and semantic indexing tasks. Overall, our study explores the feasibility of obtaining methods that scale when dealing with big data, and which enable automation of deep semantic analysis and markup of textual information from unannotated biomedical literature.

ACKNOWLEDGEMENTS

To Prof Iain Buchan and Stephen Walker for useful discussions; and to Timothy Furmston for helping with the software and e-infrastructure.

Funding: This work was supported by a grant from the European Union Seventh Framework Programme (FP7/2007-2013) for sysVASC project under grant agreement number 603288.

REFERENCES

[1] World Health Organization – Cardiovascular diseases (CVDs). Available at http://www.who.int/cardiovascular_diseases/en/. Accessed 16 June 2016.

[2] sysVASC project, http://cordis.europa.eu/project/rcn/111200_en.html. Accessed 16 June 2016.

[3] PubMed – Detailed Indexing Statistics: 1965-2015, https://www.nlm.nih.gov/bsd/index_stats_comp.html. Accessed 16 June 2016.

[4] Ely, J.W., Osheroff, J.A., Ebell, M.H., Chambliss, M.L., Vinson, D.C., Stevermer, J.J. and Pifer, E.A.: Obstacles to answering doctors' questions about patient care with evidence: qualitative study. Bmj, 324(7339), p.710 (2002).

[5] Sarker, A., Mollá, D. and Paris, C.: Automatic evidence quality prediction to support evidence-based decision making. Artificial intelligence in medicine, 64(2), pp.89-103 (2015).

[6] Hristovski, D., Dinevski, D., Kastrin, A. and Rindflesch, T.C.: Biomedical question answering using semantic relations. BMC bioinformatics, 16(1), p. 1 (2015).

[7] Tanabe, L. and Wilbur, W.J.: Tagging gene and protein names in biomedical text. Bioinformatics, 18(8), pp.1124-1132 (2002).

[8] Garten, Y., Coulet, A. and Altman, R.B.: Recent progress in automatically extracting information from the pharmacogenomic literature. Pharmacogenomics, 11(10), pp.1467-1489 (2010).

[9] Krauthammer, M. and Nenadic, G.: Term identification in the biomedical literature. Journal of biomedical informatics, 37(6), pp.512-526 (2004).
[10] Ananiadou, S., Kell, D.B. and Tsujii, J.I.: Text mining and its potential applications in systems biology. Trends in biotechnology, 24(12), pp.571-579 (2006).
[11] Federiuk, C.S., 1999. The effect of abbreviations on MEDLINE searching. Academic emergency medicine, 6(4), pp.292-296. (1999).
[12] Wren, J.D., Chang, J.T., Pustejovsky, J., Adar, E., Garner, H.R. and Altman, R.B.: Biomedical term mapping databases. Nucleic acids research, 33(suppl 1), pp.D289-D293. (2005).
[13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J.: Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, pp. 3111-3119 (2013).
[14] Mikolov, T., Chen, K., Corrado, G., & Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).
[15] Motik, B., Grau, B. C., Horrocks, I., Wu, Z., Fokoue, A., & Lutz, C.: Owl 2 web ontology language: Profiles. W3C recommendation, 27, 61 (2009).
[16] Studer, Rudi, V. Richard Benjamins, and Dieter Fensel. "Knowledge engineering: principles and methods." Data & knowledge engineering 25, no. 1, pp. 161-197 (1998).
[17] Bengio, Y., & Lee, H.: Editorial introduction to the Neural Networks special issue on Deep Learning of Representations. Neural networks: the official journal of the International Neural Network Society (2014).
[18] Landauer, T. K., & Dumais, S. T.: A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2), 211 (1997).
[19] Lund, K., & Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, & Computers, 28(2), 203-208 (1996).
[20] Kanerva, P., Kristofersson, J., & Holst, A.: Random indexing of text samples for latent semantic analysis. In Proc. of the cognitive science society (Vol. 1036). Mahwah, NJ: Erlbaum. (2000).
[21] Hofmann, T.: Probabilistic latent semantic indexing. In Proc. of ACM SIGIR conference on Research and development in information retrieval. ACM. pp. 50-57 (1999).
[22] Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent dirichlet allocation. the Journal of machine Learning research, 3, pp. 993-1022 (2003).
[23] Cohen, T., & Widdows, D.: Empirical distributional semantics: methods and biomedical applications. Journal of biomedical informatics, 42(2), pp. 390-405 (2009).
[24] Jonnalagadda, S., Leaman, R., Cohen, T., & Gonzalez, G.: A distributional semantics approach to simultaneous recognition of multiple classes of named entities. In Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg. pp. 224-235. (2010).
[25] Neelakantan, A., Shankar, J., Passos, A. and McCallum, A.: Efficient non-parametric estimation of multiple embeddings per word in vector space. EMNLP 2014: 1059-1069 (2014).
[26] Hu, B., Tang, B., Chen, Q. and Kang, L.: A novel word embedding learning model using the dissociation between nouns and verbs. Neurocomputing, 171, pp.1108-1117 (2016).
[27] Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., & Ananiadou, S.: Distributional semantics resources for biomedical text processing. In Proc. of Languages in Biology and Medicine (2013).
[28] Minarro-Giménez, J. A., Marín-Alonso, O., & Samwald, M.: Exploring the application of deep learning techniques on medical text corpora. Studies in health technology and informatics, 205, pp. 584-588 (2013).
[29] Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. and Martin, M.J.: UniProt: the universal protein knowledgebase. Nucleic acids research, 32(suppl 1), pp.D115-D119 (2004).
[30] Griffiths-Jones, S., Grocock, R.J., Van Dongen, S., Bateman, A. and Enright, A.J.: miRBase: microRNA sequences, targets and gene nomenclature. Nucleic acids research, 34(suppl 1), pp.D140-D144 (2006).
[31] Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S. and Fung, C.: HMDB: the human metabolome database. Nucleic acids research, 35(suppl 1), pp. D521-D526 (2007).
[32] OBI, http://www.obofoundry.org/ontology/obi.html
[33] PRO, http://www.obofoundry.org/ontology/pr.html
[34] SO, http://www.obofoundry.org/ontology/so.html
[35] GO, http://www.obofoundry.org/ontology/go.html
[36] ChEBI, http://www.obofoundry.org/ontology/chebi.html
[37] NCBI, http://www.obofoundry.org/ontology/ncbitaxon.html
[38] CL, http://www.obofoundry.org/ontology/cl.html
[39] UBERON, http://www.obofoundry.org/ontology/uberon.html
[40] PATO, http://www.obofoundry.org/ontology/pato.html
[41] RO, http://www.obofoundry.org/ontology/ro.html
[42] Arguello Casteleiro, M., Klein, J. and Stevens, R.: The Proteasix Ontology. Journal of biomedical semantics, 7(1) (2016).
[43] Klyne, G. and Carroll, J.J.: Resource description framework (RDF): Concepts and abstract syntax (2006).
[44] MEDLINE/PubMed XML Data Elements, https://www.nlm.nih.gov/bsd/licensee/data_elements_doc.html
[45] DCMI, http://dublincore.org/schemas/rdfs/
[46] Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C.: A neural probabilistic language model. The Journal of Machine Learning Research, 3, pp. 1137-1155 (2003).
[47] word2vec, http://code.google.com/p/word2vec/
[48] Rehurek, R. and Sojka, P.: Software framework for topic modelling with large corpora. In Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks (2010).
[49] Miles, A. and Bechhofer, S.: SKOS simple knowledge organization system reference. W3C recommendation, 18, W3C (2009).
[50] Nenadic, G., Ananiadou, S. and McNaught, J.: Enhancing automatic term recognition through recognition of variation. In Proc. of Computational Linguistics (p. 604). Association for Computational Linguistics (2004).
[51] Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, pp. 37-46 (1960).
[52] Harris, S., Seaborne, A. and Prud'hommeaux, E.: SPARQL 1.1 query language. W3C Recommendation, 21 (2013).
[53] Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A. and Tegnér, J.: Data integration in the era of omics: current and future challenges. BMC systems biology, 8(2), p.1 (2014).