Demonstration of a Literature Based Discovery System based on Ontologies, Semantic Filters and Word Embeddings for the Raynaud Disease-Fish Oil Rediscovery

Toby Reed[1,2], Vassilis Cutsuridis[1]
[1] School of Computer Science, University of Lincoln, Lincoln, LN6 7TS, UK.
[2] Streets Heaver Healthcare Computing, The Point, Weaver Rd, Lincoln LN6 3QN, UK
toby.reed@streets-heaver.com, vcutsuridis@lincoln.ac.uk

Abstract

A novel literature-based discovery system based on UMLS Ontologies, Semantic Filters, Statistics, and Word Embeddings was developed and validated against the well-established Raynaud's disease – Fish Oil discovery by mining different size and specificity corpora of Pubmed titles and abstracts. Results show an 'inverse effect' between open versus closed discovery search modes. In *open* discovery, a more general and bigger corpus (Vascular disease or Perivascular disease) produces better results than a more specific and smaller in size corpus (Raynaud disease), whereas in *closed* discovery, the exact opposite is true.

1 Introduction

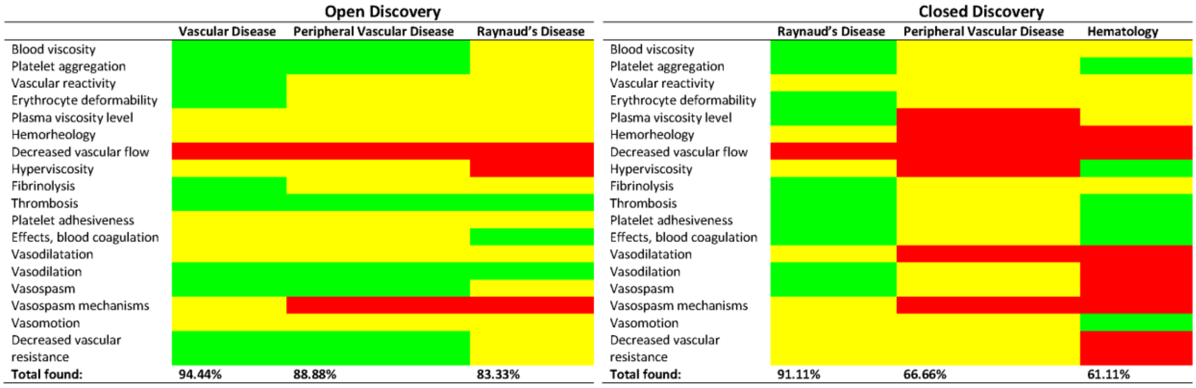
In the current COVID-19 era there is widespread demand from the pharmaceutical and healthcare industries for more work to be done in the field of reusing compounds in diseases as a method to escape some of the most expensive and time-consuming processes in drug discovery. After the most famous Sildenafil (Viagra) being repurposed from cardiovascular disease to erectile dysfunction, the use of drug repositioning has been shown to have the potential to be beneficial not only to the healthcare facilities and pharmaceutical companies, but also to the everyday consumer particularly if the process to finding and developing cures becomes cheaper, then the actual to consumer cost of treatment will likely decrease (Reed, 2020). One method for drug repositioning is through Literature Based Discovery (LBD), a powerful text mining approach that harnesses already available scientific knowledge to build bridges between seemingly unrelated islands of knowledge, such as the association of an existing drug to a novel medical condition (Reed, 2020). LBD is classified into two types: *open* and *closed* discovery. In closed discovery (also known as *hypothesis testing*), the user specifies a pair of topics (A and C) and the objective is to find any unknown, but meaningful connections (the *intermediate* (B) terms) between them. In open discovery (also known as *hypothesis generation*), the user specifies a topic of interest (C) (e.g. a disease or a drug) and the system finds a set of *intermediate* (B) terms directly related to the starting topic of interest. For each of these intermediate terms, the system reiterates the same mechanism to generate a set of *final* (A) terms.

2 Materials and Methods

A novel LBD system based on Word Embeddings, Statistics, Semantic Filters, and UMLS ontologies was developed to rediscover the Raynaud disease-Fish Oil connection (Swanson, 1986) by mining Pubmed titles and abstracts. Our system's pipeline and corpora mined to discover the Raynaud disease – Fish oil connection (Swanson, 1986) can briefly be described as follows:

1. Corpora: Different size and specificity corpora of Pubmed titles/abstracts were retrieved for each discovery type (open vs closed). *Open discovery corpora*: (i) Vascular disease, (2) Peripheral vascular disease (PVD), and (3) Raynaud disease. *Closed*

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 1–3
Patna, India, December 18 - 21, 2020. ©2019 NLP Association of India (NLPAI)



Open Discovery

	Vascular Disease	Peripheral Vascular Disease	Raynaud's Disease
Blood viscosity	🟩	🟩	🟨
Platelet aggregation	🟩	🟩	🟨
Vascular reactivity	🟩	🟨	🟨
Erythrocyte deformability	🟩	🟨	🟨
Plasma viscosity level	🟨	🟨	🟨
Hemorheology	🟨	🟨	🟨
Decreased vascular flow	🟥	🟥	🟥
Hyperviscosity	🟨	🟨	🟥
Fibrinolysis	🟩	🟨	🟨
Thrombosis	🟩	🟩	🟩
Platelet adhesiveness	🟨	🟨	🟨
Effects, blood coagulation	🟨	🟨	🟩
Vasodilatation	🟨	🟨	🟨
Vasodilation	🟩	🟩	🟩
Vasospasm	🟩	🟩	🟨
Vasospasm mechanisms	🟨	🟥	🟥
Vasomotion	🟨	🟨	🟨
Decreased vascular resistance	🟩	🟩	🟨
Total found:	**94.44%**	**88.88%**	**83.33%**

Closed Discovery

	Raynaud's Disease	Peripheral Vascular Disease	Hematology
Blood viscosity	🟩	🟨	🟨
Platelet aggregation	🟩	🟨	🟩
Vascular reactivity	🟨	🟨	🟨
Erythrocyte deformability	🟩	🟨	🟨
Plasma viscosity level	🟩	🟥	🟨
Hemorheology	🟨	🟥	🟥
Decreased vascular flow	🟥	🟥	🟥
Hyperviscosity	🟨	🟥	🟩
Fibrinolysis	🟩	🟨	🟨
Thrombosis	🟩	🟨	🟩
Platelet adhesiveness	🟩	🟨	🟩
Effects, blood coagulation	🟩	🟨	🟩
Vasodilatation	🟨	🟥	🟥
Vasodilation	🟩	🟨	🟥
Vasospasm	🟩	🟨	🟥
Vasospasm mechanisms	🟨	🟥	🟥
Vasomotion	🟨	🟨	🟩
Decreased vascular resistance	🟨	🟨	🟥
Total found:	**91.11%**	**66.66%**	**61.11%**

Figure 1: Discovered B-terms from the three open-discovery (*Left*) and closed-discovery (*Right*) corpora. In 'green' are the correctly rediscovered concepts, in 'yellow' the semantically similar discovered concepts, and in 'red' the concepts our system ought to have discovered but failed to do so.

discovery corpora: (i) Hematology, (ii) PVD, and (3) Raynaud disease. Raynaud disease is a specific type of vascular disease, but not a type of PVD, which is a sub-type of vascular disease. Raynaud disease is a sub-type of vascular disease, which involves blood, but not per se a sub-type of a hematological disease and neither is PVD.

2. Pre-processing: Each retrieved title/abstract of a scientific article was normalized to remove word variations due to capitalization. Any words with the less than three characters was also removed from further processing. All remaining words were then passed through a Natural Language Toolkit parser to generate bigrams/trigrams of each unigram based on a minimum occurrence count value.
3. A Skip-Gram Word2Vec model (Mikolov et al., 2013) was employed with some initial parameter values to generate word vectors for all words and phrases in each corpus.
4. We scanned through all generated word vectors to discover variations of the "raynaud" C-concept (e.g. Raynaud's disease, Raynaud syndrome, primary Raynaud, etc).
5. We utilised a grid search on the architecture, dimensionality, epoch, learning rate, down-sampling, context window and minimum word count parameters to find the model with the optimum performance in each corpus used.
6. Using the optimally derived Word2Vec model, we repeated STEP 4 to estimate cosine similarity of all B- or A-terms in the corpus with Raynaud variation terms from STEP 3.
7. Placed the most semantically similar terms with the closest cosine similarity, from STEP 5 into a list.
8. Mapped every term from the list via MetaMap (Aronson and Lang, 2010) to UMLS ontologies (Bodenreider, 2004). Using a semantic filter we excluded from further analysis all mapped terms which were not semantically related to the semantic types in the filter.
9. These results were then compared to previously found terms to see if our system provided acceptable results.

3 Results and Discussion

In Figure 1 results from both discovery modes. show an 'inverse effect'. In closed discovery a more specific, but smaller in size corpus (Raynaud disease) produced better results than a more general and bigger in size corpus (PVD or Hematology). On the contrary, in open discovery, a more general and bigger corpus (Vascular disease or PVD) produced better results than a more specific and smaller in size corpus (Raynaud disease). This result indicates to detect hidden relations between domain specific terms in just one-step (A-B-C), which otherwise is a multi-step process (A-B, B-C, A-C), is preferable to have more general amounts of data of the targeted domain problem to extend much further from it than to have specific data uncovering only one of these relationships (A-B or B-C). In contrast, to automatically detect words that are domain specific, is preferable to have a corpus that correctly represents the use of these specific concepts than to have more general amounts of data that encapsulate the targeted domain problem, but extend much further from it.

References

Toby S Reed. 2020. *Use of Word Embeddings in a Literature-Based Discovery System*. Master by Research Thesis, University of Lincoln, Lincoln, UK

Don R Swanson. 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine,* 31(4): 7-18

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. Efficient estimation of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS) 2013*, pages 3111-311, Lake Tahoe, Nevada.

Alan A Aronson and Francois-Michel Lang. 2010. An overview of Metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association,* 17: 229-236.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nuclei Acids Research,* 32: D267-D270.