Biomedical Text Mining for Drug Repurposing Using Word Embedding

Sruthi P. G., Shravani Sridhar, Aditya Kumar, Ahmed Bilal, R. Swathy

February 28, 2019

Drug repurposing is the process of applying an existing drug to treat a different disease than the one it was originally used for. It is becoming a common replacement to drug discovery as it has been found to be far more efficient, less costly, less time-consuming, and less risky than drug discovery. Since repurposed drugs have already passed essential safety tests during discovery, the number of stages required in drug repurposing is less.

One major challenge in drug repurposing is finding new drug-disease relationships. One approach for this is text mining, which is increasingly being used to identify and extract relationships between biological entities in literature. We focus on biomedical text mining.

Generating computational representations of linguistic units such as documents, sentences, and words is an important part of text mining. Most strategies that are used represent such units as vectors, but a disadvantage of this is that the vectors generated have too large a number of dimensions and are highly sparse. To counter this, recently a new approach has been tested, namely, word embedding, which generates relatively short numerical vectors as representations of word sense.

The word embedding algorithm approach was tested on a biomedical corpus in one study, where SVM (Support Vector Machine) was used to learn a classification model that would predict drug-disease relationships from the word vector representations and known relationships. Thus, word embedding proved to be an efficient encoding system that allows the corpus to be processed in moderate computational space and time and that also generates word vectors that are reasonably semantically equivalent to the words themselves. The final model attained a good accuracy and was successfully able to discover new drug-disease relationships; concatenating the vectors of drugs and diseases in the discovered relationships could then be used to identify candidate drugs for repurposing.

In this project, we aim to demonstrate by implementation the effectiveness of word embedding for representing senses of all words in a large amount of cancer-related biomedical literature and thus also in the discovery of novel drug-disease relationships from it for drug repurposing. The performance of the word embedding algorithm may be improved by experimenting it with feature selection and over-sampling algorithms.

References

[1] Ashburn, T.T and Thor, K. B (2004) Drug Repositioning: Identifying and Developing New Uses for Existing Drugs. Nature Reviews Drug Discovery, 3, 673-683. http://dx.doi.org/10.12688/f1000research.6653.1

[2] Luu Ngo, Duc and Yamamoto, Naoki and Anh Tran, Vu and Ngoc Giang, Nguyen and Phan, Dau and Lumbanraja, Favorisen and Kubo, Mamoru and Satou, Kenji, Application of Word Embedding to Drug Repositioning, Journal of Biomedical Science and Engineering, vol.09, pp. 7-16, 2016.

[3] Bajorath, J. (2015) Computer-Aided Drug Discovery. F1000Research, 4, 630. http://dx.doi.org/10.12688/f1000research.6653.1

[4] Emig, D., Ivliev, A., Pustovalova, O., Lancashire, L., Bureeva, S., Nikolsky, Y. and Bessarabova, M. (2013) Drug Target Prediction and Repositioning Using an Integrated Network-Based Approach. PLoS ONE, 8, e60618. http://dx.doi.org/10.1371/journal.pone.0060618

[5] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) Distributed Representations of Words and Phrases and Their Compositionality. Proceedings of NIPS. arXiv:1301.3781v3

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].Whirl-Carrillo, M., McDonagh, E.M., Hebert, J.M., Gong, L., Sangkuhl, K., Thorn, C.F., Altman, R.B. and Klein, T.E. (2012) Pharmacogenomics Knowledge for Personalized Medicine. Clinical Pharmacology Therapeutics, 92, pp. 414-417. http://dx.doi.org/10.1038/clpt.2012.96

[7] Dang, X.T., Hirose, O., Bui, D.H., Saethang, T., Tran, V.A., Nguyen, T.L.A., Le, T.T.K., Kubo, M., Yamada, Y. and Satou, K. (2013) A Novel Over-Sampling Method and Its Application to Cancer Classification from Gene Expression Data. Chem-Bio Informatics Journal, 13, pp. 19-29. http://dx.doi.org/10.1273/cbij.13.19

[8] Giorgi, J.M and Bader G.D. (2018) Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics Journal, Volume 34, Issue 23, pp. 4087–4094. https://doi.org/10.1093/bioinformatics/bty449

[9] Wei CP., Chen KA., Chen LC. (2014) Mining Biomedical Literature and Ontologies for Drug Repositioning Discovery. In: Tseng V.S., Ho T.B., Zhou ZH., Chen A.L.P., Kao HY. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2014. Lecture Notes in Computer Science, vol 8444. Springer, Cham

[10] Xue, H., Li, J., Xie, H. and Wang, Y. (2018) Review of Drug Repositioning Approaches and Resources. International Journal of Biological Sciences, 14, pp. 1232-1244. https://doi.org/10.7150/ijbs.24612