Fast neural network engine for natural science language processing: a drug-search case.

Vadim Korolev[†,‡], Artem Mitrofanov[*,†,‡], Kirill Karpov[†,‡], Valery Tkachenko[†]

†Science Data Software, LLC, 14909 Forest Landing Cir, Rockville, MD 20850, USA

‡Lomonosov Moscow State University, Department of Chemistry, Leninskie gory, 1 bld. 3, Moscow 119991, Russia

Abstract

The main advantage of modern natural language processing methods is a possibility to turn an amorphous human-readable task into a strict mathematic form. That allows to extract chemical data and insights from articles and to find new semantic relations. We propose a universal engine for processing chemical and biological texts. We successfully tested it on various use-cases and applied to a case of searching a therapeutic agent for a COVID-19 disease by analyzing PubMed archive.

Introduction

Natural language processing (NLP) is one of the driving forces of machine learning (ML) and artificial intelligence (AI) nowadays. In particular, it is of growing interest for chemical applications. An enormous quantity of published research papers[1] faced us with a necessity of its automated processing to collect, purify and organize the contained data to get new insights[2,3].

The development of NLP has been noticeably accelerated in recent years. For instance, the so-called word embeddings made it possible to translate words into high-dimensional vector space of real numbers, keeping semantic and syntactic relationships between words in a corpus[4]. Moreover, they can be used as input features for other ML models. Word embeddings have helped to significantly advance in such biomedical NLP (BioNLP) tasks as name entity recognition[5–7] (NER), information retrieval[8] (IR), question answering[9] (QA), drug-drug[10,11] and protein-protein[12] interaction extraction. The performance of these language models significantly depends on the training corpus of texts. The best results have been obtained with domain-specific corpora[13], such as abstracts[14] and full texts[15] of biomedical papers or clinical notes[16]. Unfortunately, only a few biomedical word embeddings[17–21] are publicly available to date. Moreover, most of them are provided by authors as source files, which means that their application by end-users, e.g., biochemists, requires at least basic programming skills. Online services that allow one to extract relevant information from biomedical sources are mainly based on more traditional approaches, such as the co-occurrence of terms[22].

The mentioned web-based architecture of modern solutions dictates additional restriction to a useful NLP tool: one should minimize both response time and a volume of stored data. That pushes the idea of a universal engine, able to work both on word- and text-levels. As a result, we would like to present an example of a deep-learning driven engine. The proposed solution was trained on a PubMed abstracts archive and is available to establish different NLP tasks in a reasonable time. We tested it on several classic (chemical) NLP cases and applied it to the most burning task nowadays: decision-support for a COVID-19 drug search.

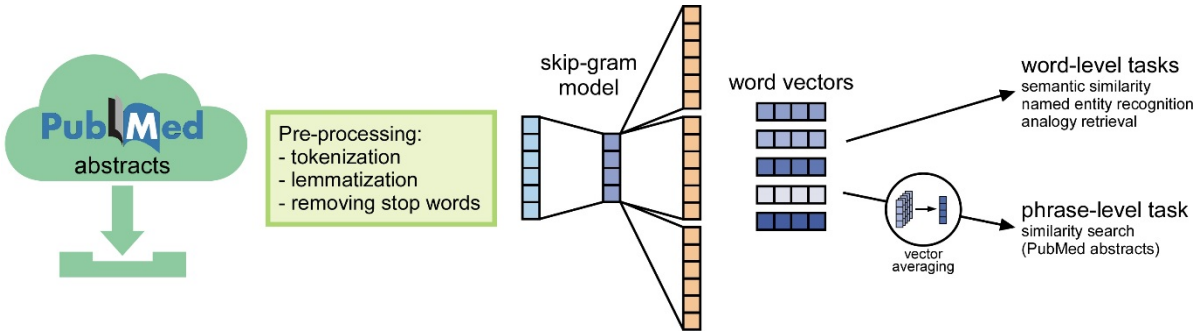


Figure 1. General workflow

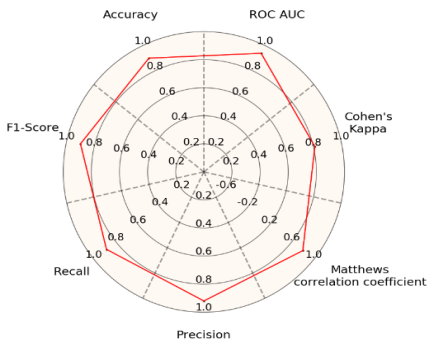
The general workflow is presented in Fig. 1. The downloaded PubMed abstracts were initially tokenized with domain-specific software, while all numbers and punctuations were converted into uniform tags. Next, pre-processed abstracts were used as input for word embedding training. At this stage, any word (token) included in the PubMed2Vec model is representable as a fixed-length vector. This representation makes it possible to operate at a word- and phrase-level (e.g. *via* word vector averaging). In addition, word vectors were also normalized to simplify the clustering process. A more detailed description of the presented pipeline is provided in the Supporting Information.

There are two main groups of validation techniques for NLP approaches. The intrinsic validation may be considered as 'down-up' or preliminary testing of the approach. We tested the engine to prove its ability to calculate reasonable similarity, to recognize the identity of term and its abbreviation, etc. More details of the first tests are available in the Supporting Information. Here we concentrated on the examples of 'up-down' validation as the solution of typical cases for NLP. The choice of the cases was based on the necessity of testing word-level processing as well as a text-level one. Also, we would like to check the performance of a highly-loaded task and try different types of ML approaches (supervised or unsupervised). As a result, we chose two cases: named entities recognition and search for similar abstracts.

Named entities recognition (NER) is one of the basic and, at the same time, the most important NLP tasks. We may define the general NER task as highlighting known entities in the text and classifying them according to some pre-defined classes. In a case of chemical information, it means identifying chemicals, drugs, biological objects in a text, i.e. compulsory for structuring knowledge and turning it into uniform data. While the highlighting part is well-studied and implemented e.g. in the PubTator[23] service, the classification part is still in the focus of investigations. As our aim was to demonstrate the capability of the engine, we used third-party PubTator software for the initial text preprocessing, while the encoding and classification tasks were performed on our pre-trained models.

For the biochemical NER we used well-known datasets based on BioNLP Shared Task series[24]. An early embeddings benchmark on several subsets[13] showed average F1-score about 0.73. The further improvement of the neutral networks architecture allowed to increase this value up to 0.80-0.86 in average, and up to 0.95 for single tasks[7,24,25]. As different research groups used different subsets from the initial database (including custom ones, available only by request), it was rather difficult to make a correct benchmark. Thus, we possessed the F1 value more than 0.80 as an acceptable value for an average biochemical NER task with modern machine learning algorithm applied.

To test the developed engine, we trained a drug/disease classifier using the dataset from [https://github.com/JHnlp/BioCreative-V-CDR-Corpus]. We used the pre-trained embeddings and gradient boosting approach[26]. The following scheme (Fig. 2) shows that the algorithm can identify diseases or chemicals with F1 score equal to 0.9 for an external test set. Thus, it proves the applicability of the proposed approach.



… anti-cancer drug paclitaxel in culture condition. … compounds including taxol. Carbohydrate-active enzymes, proteases, and secretory proteins were annotated …. … for terpene biosynthesis … pinned to taxol synthesis. (PMID:32217134)

Figure 2. The metrics (left) and example (right) of chemicals (green) / diseases (blue) classification.

The second test case was a search of texts similar to the target one. While the previous example was devoted to word processing, the second tested the proposed approach on a text-level. Here the processing time appears along with the processing quality. And the question of search time is a choice of processing algorithms as well as sensible data storage. We had to split the embedded data into separate clusters to reduce both loading and processing time. Thus, the case involves the texts clustering as well as similarity search.

The general scheme is presented in Figure 3. We used a set of normalized abstract vectors to pre-train 2000 clusters with semantic similar abstracts. Each cluster was described by a single initial centroid vector. A preliminary search stage looked for 20 clusters, that were the most similar to the target one. The second stage processed the chosen abstracts in a parallel mode to identify the most similar abstracts. The described approach allowed to reduce the search time for nearly 12 seconds comparing the hours for the basic sequential grid search.

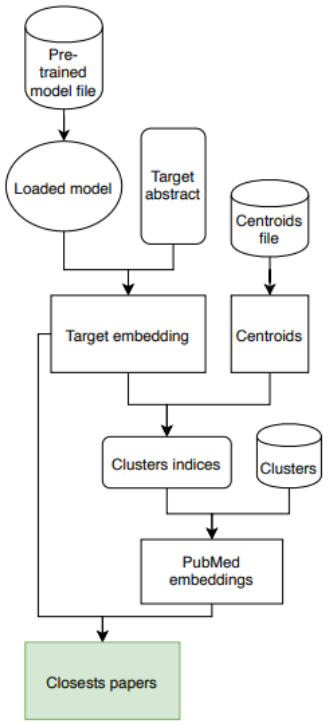


Figure 3. The general workflow of an algorithm searching for similar abstracts.

As an example of the system's work, we performed a search for articles similar to an article with PMID: *3966546*. A search was performed according to the above scheme within our system, and, for comparison, a similar search was performed using the search engine of PubMed itself. The search results are presented in Figure 4. The top article for the original PubMed search engine also appeared in our top list (green dot). The target publication was devoted to the investigations of the Guinea pig embryo, namely, the change in its physicochemical composition during embryonic development. PubMed search results include publications on similar topics, all of which relate directly to Guinean pigs. The search results obtained through our search include publications directly on the development of Guinea pigs, as well as articles on similar topics that are not directly related to a specific animal species or that describe other species. We can also mention, the original frequency-based search looks for a compact topic, while the embedding-based allows to perform broader search.

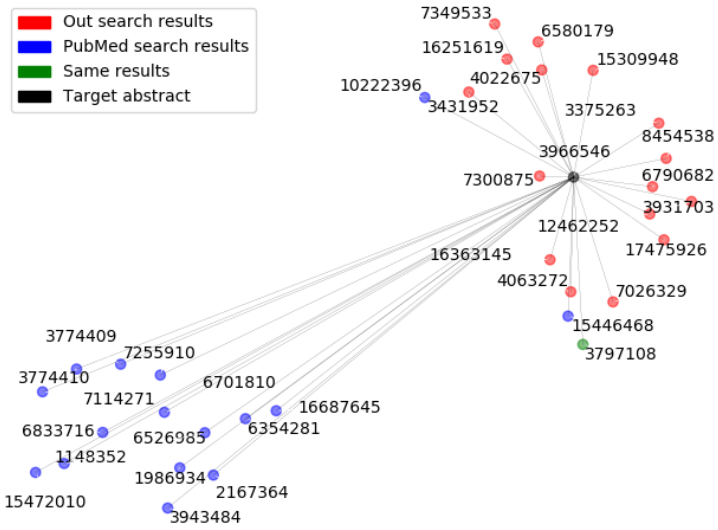


Figure 4. Comparison of the embedding-base similarity search (this article, red dots) with frequency-based PubMed search (blue dots)

Finally, we applied the proposed NLP architecture to a drug search task. Word embeddings based on neural network models have been proven to capture properly semantic relationships[4,27]. In particular, word embedding models allow retrieving analogy terms. Given a pair of related terms a and b, we can determine unknown analogy word d for term c by applying the simple semantic operation:

$$w_d = w_c - w_a + w_b$$

where w_a, w_b, w_c, and w_d are word embedding vectors for corresponding words. The calculated vector w_d usually does not exactly match any of the words from the vocabulary. Thus, it worth considering several nearest words in the vector space in terms of cosine similarity. By embracing this concept, we considered resolving drug-disease relationships for drug repurposing.

We extracted the described relation ($w_c - w_a$ vectors) for several drug/disease pairs and tested the approach (Fig. 5). In all the cases the algorithm proposed the required drugs in the output. It means, the algorithm compressed the impossible task of manual reading of thousands of articles with the mentioned terms inside, to a simple task of choosing from ten options.

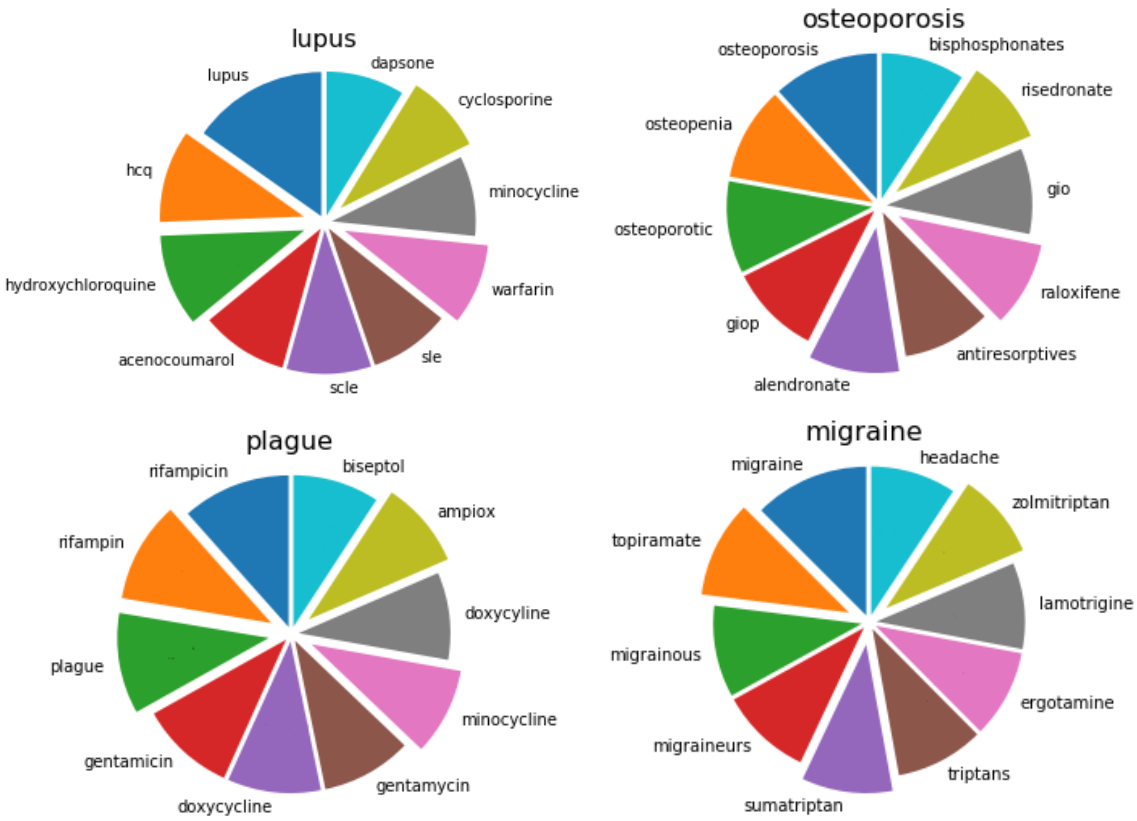


Figure 5. The proposed 'drugs' for a set of chosen diseases. The highlighted sectors represent real drugs, relevant to the diseases.

The outbreak of the novel coronavirus disease in 2019 (COVID-19) attracted considerable attention from the global scientific community. Significant efforts are focused on the search for therapeutic agents and the development of preventive vaccines[28]. The former well formalizable task can be considered using the PubMed2Vec model. Given the already known pairs disease – therapeutic agent, where the disease is related to COVID-19, such as Middle East respiratory syndrome (MERS), we can obtain the alleged therapeutic agents for COVID-19 using the semantic arithmetic described above. In addition, we considered disease – therapeutic agent pairs, where the therapeutic agent is an existing drug with therapeutic potentials for COVID-19, according to drug repurposing conception[28]. The used pairs of terms and the most likely candidates are presented in Table 2 and Table 3, respectively.

Table 2. Drug-disease pairs using to define promising candidates for COVID-19.

drug	disease
baricitinib	rheumatoid arthritis (**RA**)
remdesivir	Ebola virus infection (**EVD**)
galidesivir	Ebola virus infection (**EVD**)
chloroquine	malaria
Arbidol	influenza
Interferon beta-1b (**interferon-beta-1b**)	Middle East respiratory syndrome (**MERS**)
lopinavir	Middle East respiratory syndrome (**MERS**)

ritonavir	Middle East respiratory syndrome (**MERS**)

Table 3. Most promising candidates (as therapeutic agents for COVID-19) obtained via semantic arithmetic.

drug	cosine similarity
atazanavir	0.6949
darunavir	0.6879
indinavir	0.6822
saquinavir	0.6718
efavirenz	0.6584
nelfinavir	0.6567
fosamprenavir	0.6455
amprenavir	0.6431
raltegravir	0.6382
solumedrol	0.5887
betamethazone	0.5794
betaferon	0.5751
simeprevir	0.5378
favipiravir	0.5080
t-1105	0.5032

Thus, we have demonstrated how the presented PubMed2Vec engine can act in two modalities (word- and sentence-level), helping in named entity recognition, similarity search, and resolving drug-disease relationships. Now we are working on the graphical interface, but we hope even the current results will be helpful.

References

(1) Khare, R.; Leaman, R.; Lu, Z. Accessing Biomedical Literature in the Current Information Landscape. In *Biomedical Literature Mining*; Springer, 2014; pp 11–31.

(2) Cohen, K. B.; Hunter, L. Getting Started in Text Mining. *PLoS Comput. Biol.* **2008**, *4* (1).

(3) Rzhetsky, A.; Seringhaus, M.; Gerstein, M. B. Getting Started in Text Mining: Part Two. *PLoS Comput. Biol.* **2009**, *5* (7).

(4) Mikolov, T.; Yih, W.; Zweig, G. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*; 2013; pp 746–751.

(5) Tang, B.; Cao, H.; Wang, X.; Chen, Q.; Xu, H. Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks. *Biomed Res. Int.* **2014**, *2014*.

(6) Habibi, M.; Weber, L.; Neves, M.; Wiegandt, D. L.; Leser, U. Deep Learning with Word Embeddings Improves Biomedical Named Entity Recognition. *Bioinformatics* **2017**, *33* (14), i37--i48.

(7) Yoon, W.; So, C. H.; Lee, J.; Kang, J. CollaboNet: Collaboration of Deep Neural Networks for Biomedical Named Entity Recognition. *BMC Bioinformatics* **2019**, *20* (10), 249.

(8) Mohan, S.; Fiorini, N.; Kim, S.; Lu, Z. A Fast Deep Learning Model for Textual Relevance in Biomedical Information Retrieval. In *Proceedings of the 2018 World Wide Web Conference*; 2018; pp 77–86.

(9) Wiese, G.; Weissenborn, D.; Neves, M. Neural Domain Adaptation for Biomedical Question Answering. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*; 2017; pp 281–289.

(10) Liu, S.; Tang, B.; Chen, Q.; Wang, X. Drug-Drug Interaction Extraction via Convolutional Neural Networks. *Comput. Math. Methods Med.* **2016**, *2016*.

(11) Zhang, Y.; Zheng, W.; Lin, H.; Wang, J.; Yang, Z.; Dumontier, M. Drug--Drug Interaction Extraction via Hierarchical RNNs on Sequence and Shortest Dependency Paths. *Bioinformatics* **2018**, *34* (5), 828–835.

(12) Jiang, Z.; Li, L.; Huang, D. A General Protein-Protein Interaction Extraction Architecture Based on Word Representation and Feature Selection. *Int. J. Data Min. Bioinform.* **2016**, *14* (3), 276–291.

(13) Wang, Y.; Liu, S.; Afzal, N.; Rastegar-Mojarad, M.; Wang, L.; Shen, F.; Kingsbury, P.; Liu, H. A Comparison of Word Embeddings for the Biomedical Natural Language Processing. *J. Biomed. Inform.* **2018**, *87*, 12–20.

(14) https://www.ncbi.nlm.nih.gov/pubmed.

(15) https://www.ncbi.nlm.nih.gov/pmc.

(16) Johnson, A. E. W.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; Mark, R. G. MIMIC-III, a Freely Accessible Critical Care Database. *Sci. data* **2016**, *3*, 160035.

(17) Björne, J.; Van Landeghem, S.; Pyysalo, S.; Ohta, T.; Ginter, F.; de Peer, Y.; Ananiadou, S.; Salakoski, T. PubMed-Scale Event Extraction for Post-Translational Modifications, Epigenetics and Protein Structural Relations. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*; 2012; pp 82–90.

(18) Moen, S.; Ananiadou, T. S. S. Distributional Semantics Resources for Biomedical Text Processing. *Proc. LBM* **2013**, 39–44.

(19) Chiu, B.; Crichton, G.; Korhonen, A.; Pyysalo, S. How to Train Good Word Embeddings for Biomedical NLP. In *Proceedings of the 15th workshop on biomedical natural language processing*; 2016; pp 166–174.

(20) Zhang, Y.; Chen, Q.; Yang, Z.; Lin, H.; Lu, Z. BioWordVec, Improving Biomedical Word Embeddings with Subword Information and MeSH. *Sci. data* **2019**, *6* (1), 1–9.

(21) Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* **2020**, *36* (4), 1234–1240.

(22) Capuzzi, S. J.; Thornton, T. E.; Liu, K.; Baker, N.; Lam, W. I.; O'Banion, C. P.; Muratov, E. N.; Pozefsky, D.; Tropsha, A. Chemotext: A Publicly Available Web Server for Mining Drug--Target--Disease Relationships in PubMed. *J. Chem. Inf. Model.* **2018**, *58* (2), 212–218.

(23) Wei, C.-H.; Allot, A.; Leaman, R.; Lu, Z. PubTator Central: Automated Concept Annotation for

Biomedical Full Text Articles. *Nucleic Acids Res.* **2019**, *47* (W1), W587–W593. https://doi.org/10.1093/nar/gkz389.

(24) Crichton, G.; Pyysalo, S.; Chiu, B.; Korhonen, A. A Neural Network Multi-Task Learning Approach to Biomedical Named Entity Recognition. *BMC Bioinformatics* **2017**, *18* (1), 368. https://doi.org/10.1186/s12859-017-1776-8.

(25) Cho, H.; Lee, H. Biomedical Named Entity Recognition Using Deep Neural Networks with Contextual Information. *BMC Bioinformatics* **2019**, *20* (1), 735. https://doi.org/10.1186/s12859-019-3321-4.

(26) Chen, T.; Guestrin, C. XGBoost : A Scalable Tree Boosting System. *arXiv:1204.1234* **2016**.

(27) Liu, F.; Chen, J.; Jagannatha, A.; Yu, H. Learning for Biomedical Information Extraction: Methodological Review of Recent Advances. *arXiv Prepr. arXiv1606.07993* **2016**.

(28) Liu, C.; Zhou, Q.; Li, Y.; Garner, L. V; Watkins, S. P.; Carter, L. J.; Smoot, J.; Gregg, A. C.; Daniels, A. D.; Jervey, S.; et al. Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases. *ACS Central Science*. 2020, pp 315–331.