Linking Semantic Fingerprints of Literature – from Simple Neural Embeddings Towards Contextualized Pharmaceutical Networks

Janus Wawrzinek[1][0000-0002-8733-2037], José María González Pinto[1][0000-0002-2908-3466] and Wolf-Tilo Balke[1][0000-0002-5443-1215]

[1] IFIS TU-Braunschweig, Mühlenpfordstrasse 23, 38106 Braunschweig, Germany
{wawrzinek, pinto, balke}@ifis.cs.tu-bs.de

Abstract. The exponential growth of publications in medical digital libraries requires new access paths that go beyond term-based searches, as these increasingly lead to thousands of results. An effective approach for this problem is to extract important pharmaceutical entities and their relations to each other in order to reveal the embedded knowledge in digital libraries. State-of-the-art approaches in the field of neural-language models (NLMs) enable progress in learning and predicting such relations in terms of semantic quality, scalability, and performance and already now make them valuable for important research tasks such as hypothesis generation. However, in the field of pharmacy a simple list of (predicted) associations is often challenging to interpret because, between typical pharmaceutical entities, such as active substances, diseases, and genes, complex associations will exist. A *contextualized network* of pharmaceutical entities can support the exploration of these associations and will help to assess and interpret predicted relationships. On the other hand, the prerequisite for building meaningful entity networks is an answer to the question: *When is an NLM-learned entity relation meaningful?* In this paper, we investigate this question for important pharmaceutical entity relations in the form of drug-disease associations (DDAs). To do so, we present a *new methodology* to determine entity-specific thresholds for the existence of associations. Such entity-specific thresholds open-up the possibility of automatically constructing (meaningful) embedded pharmaceutical networks, which can then be used to explore and to explain learned relationships between pharmaceutical entities.

Keywords: digital libraries, information extraction, neural embeddings.

1 Introduction

The increased quantity of publications in Digital Libraries challenges users that need to collect, understand, and integrate a broad range of relevant and related knowledge to design new hypotheses. As a good example, consider entity-based searches in scientific digital libraries. The keyword search for the disease "*Ovarian Neoplasms*" in the PubMed medical library leads to a result set of over 96,000 documents. One of the most effective ways to address this problem is to extract important pharmaceutical entities such as active ingredients, diseases, and genes plus their mutual relationships to reveal the embedded knowledge contained in such massive collections. In this context state of the art Neural Language Models (NLMs) can be used not only to extract known relationships, but also to predict new relationships [6] such as Drug-Disease Associations (DDAs). A DDA exists when a) a drug helps against a particular disease (cures, prevents, alleviates) or b) a drug induces a disease in the sense of a side effect [3]. Such DDAs are important to explore, for instance, because they are possible candidates for drug-repurposing [5]. Indeed, pharmaceutical research often focuses on well-known and proven active substances against other diseases, as this generally leads to lower risk in terms of adverse side effects. The question is how can we support users in the exploration and assessment of predicted Drug-Disease-Associations? In this context, network views constructed using *k*-nearest-neighbour (kNN) sets can help users to understand complex entity associations [1] better. For instance, extracting *k*-nearest-disease sets for each drug and using them for building a drug-disease network could give users an overview of all embedded drug-disease associations learned by the NLM and thus may support users in formulating new hypothesises for predicted associations. *How to build such k*-nearest-neighbour *sets?* Using similarity thresholds for the extraction of the *kNN* sets is error-prone, because threshold-based approaches are hard to estimate and to interpret [6, 8, 9] and may vary for different entities, time-periods, and corpus sizes [6, 10]. Thus, as a prerequisite for building meaningful drug-disease networks, we have first to answer the question: *How to find a meaningful k for each embedded drug-entity?*

In this paper, we present a *new literature-based methodology* which enables us to predict how many *k*-Nearest-Disease-Neighbors (*k*-NDNs) we should extract per drug-entity from the embedding space, and this stays relatively stable for different corpora sizes, time-periods, and document lengths. Even if our research focuses on a specific task, we believe that our investigations lead to a general better understanding of state-of-the-art Neural Language Models like Word2Vec. This is a short version of our extended study. Thus for the interested reader, we have our full technical report available [10]. The paper is organized as follows: Section 2 revisits related work accompanied by our extensive investigation of entity-specific thresholds in section 3. We close with conclusions in section 4.

2 Related Work

Research in the field of digital libraries has been dealing with semantically meaningful similarities for entities and their relations for a long time. For the investigation of semantic relations between words, Neural Language Models (NLMs) like Word2Vec are currently the state-of-the-art approaches [4]. Word2Vec models can also be divided into the categories Continuous Bags of Words (CBOW) and the Skip-gram model. The difference here is how word-embeddings are learned; for example, CBOW tries to predict the matching word with a word context, while Skip-gram tries to predict a word context with a word. An important aspect here is that the Skip-gram model is better suited for semantic tasks [4]. Therefore, in our investigation, we will rely on the Word2Vec Skip-

Gram model implementation from the open source Deep-Learning-for-Java[1] library. With the increasing popularity of predictive models, interest in the study of the semantic meaning of distance in high-dimensional spaces is growing. Elekes et al. [8] investigated the influence of hyperparameters and document corpus size to the similarity of word pairs. In their investigation, they compare word pair distances in the embedding space with a WordNet Similarity. They also point out that similarity of words in natural language is blurred and therefore problematic to measure. In contrast to natural language, the word pairs we are investigating feature rather a binary than a blurred relation to each other (a drug *x* affects disease *y* or not [3]). In our investigation, we measure the quality of this binary relation in the word embedding space. The next question is whether meaningful similarity thresholds exist for semantically associated word pairs. However, similarity thresholds are difficult to calculate and can vary for different models and corpora [6, 8]. Therefore, we do not determine the thresholds with a similarity comparison but use corpus information in order to determine how many neighbors we should extract from the embedding space.

3 Experimental Investigation

We will first describe our pharmaceutical text corpus and experimental set-up decisions. Afterward, we investigate for various time-periods the general correlation between embedded DDAs and corpora-statistics (section 3.1). We continue our correlation analysis involving predicted DDAs (section 3.2) and propose a method for entity-specific thresholds. Finally, we compare the efficiency of our entity-specific k-NDNs thresholds (section 3.3) with two conventional approaches (fixed k-NDNs, Similarity-Threshold). As shown in [6] on the one hand DDA predictions are possible, but on the other hand, the majority of existing DDAs cannot be extracted from the embedding space. Therefore, we focus on precision as the main quality-measure.

Experimental Setup.

Evaluation corpus. PubMed[2] is with more than 29 million document citations, the largest and most comprehensive digital library in the biomedical field. Since full-text access is not available for the most publications, we used only abstracts for our evaluation corpus (More details about the corpus can be found here [10]).

Time Period Evaluation Corpora. In order to calculate the change in different time periods as well as on different corpus sizes, we divide our evaluation corpus into four corpora: **1900-1988, 1900-1998, 1900-2008, 1900-2018**. Each corpus contains only the documents for the respective time period.

Query Entities. As query entities for the evaluation, we randomly selected 350 drugs from the *DrugBank*[3] collection, which is a 10% sample of all approved drugs. Thus, our final document set for evaluation contains ~2.5 million abstracts for 350 drugs.

[1] https://deeplearning4j.org/

[2] https://www.ncbi.nlm.nih.gov/pubmed/

[3] https://www.drugbank.ca/

Experiment implementation and parameter settings.

1. *Text Preprocessing*. Stop-word removal and stemming were performed using a *Lucene*[4] index. For stemming we used Lucene's *Porter Stemmer* implementation. We considered all words contained in more than half of the indexed abstracts as stop-words. Here we made sure that the drug and disease identifiers were not affected.
2. *Word Embeddings*. After pre-processing, word embeddings were created with DeepLearning4J's *Word2Vec*[5] implementation.
3. *Hyperparameter-Tuning*. Larger window-sizes may affect association learning quality [7]. Based on the extended experiments described in our technical report [10], we will investigate the effect of a default window-size setting (w = 5) as well as with a higher window-size (w = 50) in our correlation-experiments.
4. *Similarity-Measure*. As the similarity measure between the drug/disease embeddings, we choose cosine similarity in all experiments. A value of 1 between two vectors means a perfect similarity (vectors match perfectly) and the value 0 means a maximum dissimilarity (vectors are orthogonal to each other).

3.1 Relationship between Corpus-Data and Existing-DDAs

In our first experiment we clarify if there exists always, and to which extent, a general correlation between *corpus data* and the number of *existing* DDAs (DDA appears in at least three abstracts) to be extracted from the embedded space. As *corpus data,* we use the following data sets:

1. ***Document-Count (Doc-Count)***: We count all abstracts per drug entity in which the drug is present.
2. ***DDA-Count***: For each drug, we count the number of all DDAs existing in the abstracts. Here, we count only DDAs which are present in at least three abstracts.

Evaluation Implementation. In our experiment, we extract for each drug-entity 20-NDNs from the embedding space and measure the AVG-Precision of the *existing* DDAs occurring in literature. For our experiment, we sorted the two data set lists per drug-entity in ascending order and divided them into ten percentile groups. We have performed this grouping in order to ensure a better overview and comparability in the presentation. Therefore, for example, the first group contains the first 10% of all drug-entities with the lowest document counts or DDA counts, and the 10th group the 10% of drug-entities with the highest counts. As window size, we have chosen w = 50 and set the number of dimensions to 200. We experiment with the four corpora already introduced: 1900-1988, 1900-1998, 1900-2008, 1900-2018.

What we can see is that with increasing amounts of data (higher group) as well as in all corpora, the precision increases (Figure 1a, 1b). Therefore, as expected, the more

[4] https://lucene.apache.org/

[5] https://deeplearning4j.org/word2vec

data available per drug-entity, the more *existing*-DDAs can probably be extracted. For the corpora 1900-2008 and 1900-2018 the values are almost the same.

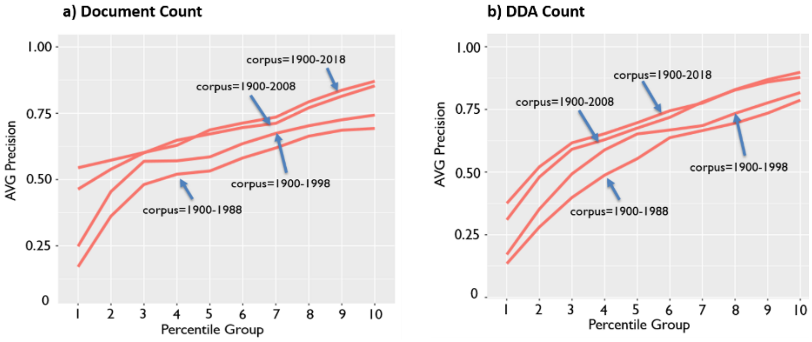

Fig. 1. AVG precision values reached for different corpora and percentile groups with Document-Count (a) and DDA-Count (b).

Thus, from a certain amount of data, there seems to be no substantial improvement in precision. The curves for the DDA-Count (Figure 1b) seem to be more similar in shape and course than the Document-Count curves (Figure 1a). This could indicate that DDA-Count data correlation is more stable over periods and corpora. This experiment has shown us that a certain correlation between data volume and individual entity probably exists. We will investigate this possible correlation, its expression, and stability in more detail in our subsequent experiments to determine entity thresholds.

3.2 Relationship between Corpus-Data and Predicted DDAs

Given our findings of the previous section, here we investigate how we can find possible thresholds if DDA predictions are involved and that we are precision-oriented. In other words, how do we choose a *meaningful* precision value for our further investigations? If we set the precision p too high (p > 95%) we have fewer chances of making predictions. If we set the precision too low (p < 50), the majority of DDAs are not meaningful at a current point in time. In order to choose a meaningful precision value, we have made a ***majority assumption***: If the majority (here in AVG-p ~67%) of the DDAs exists - each DDA occurs in at least three abstracts - then there is a probability that the remaining minority is also meaningful and represents DDA predictions. Based on this majority assumption and the determination of a *meaningful* precision value, we customize our correlation analysis: What is the number k, of NDNs we need to extract to achieve AVG (*existing)*-DDA precision of ~67%? Moreover, how strong does k correlate with the amount of data d, where d = Document-Count or d = DDA-Count? If there is a strong linear correlation, we can, for example, train a regression model and for a given amount of data d of a drug x, to determine the number k of NDNs leading probably to the desired DDA-precision of ~67%.

Evaluation Implementation. To measure changes in the correlation over time, we train the models again on the four corpora: 1900-1988, 1900-1998, 1900-2008, and

1900-2018. Besides, we train the models with 200 dimensions, with the best window size (*w*=50) and the default setting (*w*=5) to determine the influence of the window size on the correlations. For all combinations between Document-/DDA-Count and number of NDNs to be selected to achieve ~67% precision, we determine the Pearson Correlation Coefficient (PCC).

Table 1. PCCs for different window sizes and data counts. Best values in bold (SD=Standard Deviation).

	DDA-Count		Doc-Count	
Corpus	(w=50)	(w=5)	(w=50)	(w=5)
1900-2018	**0.763**	0.530	0.590	0.453
1900-2008	**0.732**	0.634	0.509	0.448
1900-1998	**0.726**	0.640	0.444	0.386
1900-1988	**0.756**	0.725	0.439	0.434
Mean:	**0.744**	0.632	0.495	0.431
SD:	**0.018**	0.079	0.070	0.030

Result interpretation. One of the more interesting and rather counterintuitive results (Table 1) of our experiment is that the correlation can decrease with increasing data volume. This happens when DDA-Count is used with the combination *w*=5. The correlation decreases continuously from 0.725 (corpus 1900-1988) to 0.530 (corpus 1900-2018). *How can this result be explained?* The length of the abstracts in PubMed has grown continuously over the last four decades [2]. This, of course, affects the training, because with increasing text length the probability for more DDAs within an abstract increases. This, in turn, means that smaller entity contexts combined with longer texts potentially contain fewer DDAs, and thus the correlation decreases. This thesis is supported by the fact that this effect does not occur with larger entity contexts (DDA count with w=50). This leads to two insights for larger entity contexts in combination with DDA-Count: a) The correlation is always strong (AVG PCC = 0.744) and b) is hardly influenced by the size of the corpus as well as by the abstract length (SD= 0.018). Thus, there is a chance that we can train a regression model based on the DDA-Count and at a window size of 50 and this independent of corpus size and document length. Besides the described properties, this finding can be useful in the historical analysis of drugs. The results for Document-Count (Table 1) confirm the hypothesis that with increasing data also learning quality improves, and this leads to increasing correlation values. On the other hand, only weak AVG correlation can be achieved with Document-Count (AVG PCC < 0.5). In general, the correlation with a DDA-Count is always stronger compared to Document-Count. Therefore, it would be probably advantageous to use only documents with contained DDAs for training.

3.3 Learning Dynamic Thresholds for k-NDNs

Now it has to be tested if a predicted *k*-NDN-threshold approach is more efficient than, e.g., using a fixed *k* or determining a specific similarity value. How do we compare the performance of these thresholds? A comparison with precision values is less meaningful, because the fewer NDNs we choose per drug entity, the higher the precision, the lower the recall and probably fewer DDA predictions are possible. Efficient in our case means that with a given precision (~67%) we can achieve a higher recall.

Experimental Evaluation. We investigate the efficiency of our thresholds again for different periods using the four corpora already introduced. First, we train the embedding models with a window size of *w*=50 and a dimension size of *d*=200 with the respective corpora. Then we train a regression model to determine the *k*-NDNs. For this, we use the DDA-Count and the number of k-NDNs which lead to an AVG DDA precision of 67%. We perform 10-fold cross-validation and train with 90% of the data and test with 10% of the remaining data. The recall is always determined in the current period. Therefore, for example, in 1900-1988, we can determine 8% of all DDAs existing at that time (Table 2).

Table 2. Recall values for the different approaches and corpora. Precision is fixed in all approaches to ~67%. Best values in bold.

Time periods	Predicted Threshold	Fixed k-NDNs	Similarity Threshold
1900-2018	**0.07**	0.03	0.04
1900-2008	**0.07**	0.04	0.06
1900-1998	**0.07**	0.03	0.04
1900-1988	**0.08**	0.02	0.04

Comparative approaches. We perform a comparison with two conventional approaches (fixed *k*-NDNs, Similarity-Threshold). First, we determine the fixed number of k-NDNs, which lead to a DDA AVG precision of ~67%. As an example, 6-NDNs must be extracted from the embedding space per drug-entity in the period 1900-1988 in order to achieve a DDA AVG precision of 67%. For further comparison, we determine a similarity value which leads to an AVG DDA precision of ~67%. Where a Similarity value of 0 means maximum inequality and a value of 1 means maximum equality. For example, the Similarity value in the period 1900-1998 is ~0.38 to reach an AVG precision of ~67%.

As presented in Table 2, our approach leads to improved recall values. On average, the recall increases by about 61% compared to Similarity-Threshold and 142% compared to a fixed-k-NDNs approach. Thus, we can extract more DDAs from the embedded space. The recall is *relatively* stable for the different time-periods and corpora. Thus, we can rely on certain value-stability.

4 Conclusions

State-of-the-art approaches in the field of Neural Language Models (NLMs) enable progress in learning and predicting entity associations in terms of semantic quality, scalability, and performance. In this context, network views constructed using *k*-nearest-neighbour sets can help users to understand better complex entity associations [1] contained in literature and thus help to explain *why* an NLM has learned or predicted a certain entity association. On the other hand, learning quality varies per entity and (default) approaches like fixed-*k*-NN or Similarity-Thresholds are less meaningful for the extraction of *k*-nearest-neighbour sets from the embedded space. In this paper, we presented a novel literature based method to learn entity-specific thresholds for pharmaceutical entities. This enables us to become independent from the difficult to determine *k*-NN or Similarity-Thresholds and opens up the possibility to develop a more meaningful contextualized drug-disease network using *k*-nearest-neighbour sets.

References

1. Greene, D., & Cunningham, P. (2013, May). Producing a unified graph representation from multiple social network views. In *Proceedings of the 5th annual ACM web science conference* (pp. 118-121). ACM.
2. MEDLINE®/PubMed® Data Element (Field) Descriptions. (U.S. National Library of Medicine). Retrieved April 4, 2019, from https://www.nlm.nih.gov/bsd/mms/medlineelements.html#ab
3. Zhang, W., Yue, X., Chen, Y., Lin, W., Li, B., Liu, F., & Li, X. (2017, November). Predicting drug-disease associations based on the known association bipartite network. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 503-509). IEEE.
4. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
5. Dudley, J. T., Deshpande, T., & Butte, A. J. (2011). Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics*, *12*(4), 303-311.
6. Wawrzinek, J., & Balke, W. T. (2018, November). Measuring the Semantic World–How to Map Meaning to High-Dimensional Entity Clusters in PubMed? In *International Conference on Asian Digital Libraries* (pp. 15-27). Springer, Cham.
7. Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665-695.
8. Elekes, Á., Schäler, M., & Böhm, K. (2017, June). On the Various Semantics of Similarity in Word Embedding Models. In *Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on* (pp. 1-10). IEEE.
9. Al-Natsheh H.T., Martinet L., Muhlenbach F., Rico F., Zighed D.A. (2018) Metadata Enrichment of Multi-disciplinary Digital Library: A Semantic-Based Approach. In: Méndez E., Crestani F., Ribeiro C., David G., Lopes J. (eds) Digital Libraries for Open Knowledge. TPDL 2018. Lecture Notes in Computer Science, vol 11057. Springer.
10. Wawrzinek, J., Pinto, J. M., & Balke, W. T. (2019). Linking Semantic Fingerprints of Literature – from Simple Neural Embeddings Towards Contextualized Pharmaceutical Networks (Supplement). Retrieved June 18, 2019, from http://www.ifis.cs.tu-bs.de/sites/default/files/wawrzinek-pinto-balke-Technical-Report.pdf