Measuring the Semantic World – How to Map Meaning to High-Dimensional Entity Clusters in PubMed?

Janus Wawrzinek[1[0000-0002-8733-2037]] and Wolf-Tilo Balke[1[0000-0002-5443-1215]]

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

Abstract. The exponential increase of scientific publications in the medical field urgently calls for innovative access paths beyond the limits of a term-based search. As an example, the search term "diabetes" leads to a result of over 600,000 publications in the medical digital library PubMed. In such cases, the automatic extraction of semantic relations between important entities like active substances, diseases, and genes can help to reveal entity-relationships and thus allow simplified access to the knowledge embedded in digital libraries. On the other hand, for semantic-relation tasks distributional embedding models based on neural networks promise considerable progress in terms of accuracy, performance and scalability. Yet, despite the recent successes of neural networks in this field, questions arise related to their non-deterministic nature: Are the semantic relations meaningful, and perhaps even new and unknown entity-relationships? In this paper, we address this question by measuring the associations between important pharmaceutical entities such as *active substances (drugs)* and *diseases* in high-dimensional embedded space. In our investigation, we show that while on one hand only few of the contextualized associations directly correlate with spatial distance, on the other hand we have discovered their potential for predicting new associations, which makes the method suitable as a new, literature-based technique for important practical tasks like e.g., drug repurposing.

Keywords: digital libraries, information extraction, deep neuronal embeddings.

1 Introduction

In digital libraries the increasing information flood requires new and innovative access paths that go beyond simple term-based searches. This is of particular interest in the scientific field where the number of publications is growing exponentially [17] and access to knowledge is getting increasingly difficult, such as to (a) *medical entities* like active substances, diseases, or genes and (b) *their relations*. However, these entities and their relations play a central role in the exploration and understanding of entity relationships [2]. Extracting entity relations automatically is therefore of particular interest, because it bears the potential for new insights and numerous innovative applications in important medical research areas such as the discovery of new drug-disease associations (DDAs) needed, e.g., for drug repurposing [14]. Previous work has recognized

this trend and focuses on the recognition of these pharmaceutical entities and their relationships [3, 6, 9]. *What is a DDA?* A DDA is in general an effect of drug *x* on disease *y* [3], which means a) an active substance helps (cures, prevents, alleviates) a certain disease or b) an active substance causes/triggers a disease in the sense of a side effect. *Why are DDAs of interest?* DDAs are considered as potential candidates for drug repurposing. Pharmaceutical research attempts to use well-known and well-proven active substances against other diseases. This generally leads to a lower risk (in the sense of well-known side effects [4]) and significantly lower costs [4, 6]. Based on the interests mentioned above, numerous computer-based methods were developed to derive DDAs from text corpora as well as from specialized databases [9]. The similarity between active substances and diseases forms the basis here, and numerous popular methods exists for calculating a similarity between these pharmaceutical entities such as chemical (sub) structure similarity [8] or network-based similarity [4].

Newer approaches attempt to derive an intrinsic connection between pharmaceutical entities with a linguistic, context-based similarity [7, 10, 11]. The basic idea is the distributional hypothesis: a similar linguistic word-context indicates a similar meaning (or properties) of the entities contained in texts. In this kind of entity-contextualization the currently popular distributional semantic models (also embedding models or neural language models) play a major role because they enable an efficient way for learning semantic word-similarities in huge corpora. However, non-deterministic word-embedding models like Word2Vec are on one hand popular models for semantic, as well as analogy tasks, but on the other hand their properties are not fully investigated [15].

In this context, we address the following research questions: (Q1) Is a meaningful DDA also represented in the word embedding space and can we measure it in terms of a linguistic distance? (Q2) How can this be measured, evaluated, and what are meaningful baselines and datasets? (Q3) Since a word-context is learned on the basis of millions of publications, is new knowledge discovered with this kind of contextualization? (Q4) Are even DDA predictions possible with such models, and if so, how high is the respective *predictive factor*?

In this paper we answer all these questions and follow our use case of pharmaceutical entities drug/disease and their relations (DDAs). We evaluate the embedded entities both with manually curated data from specialized pharmaceutical databases as well as text-mining approaches. In addition, we carry out a retrospective analysis, which shows that low-distance relations (which previously did not explicitly occur in documents) will actually occur in future publications with a high probability. This indicates that a future relation already exists at an early stage in embedded space which can also help us to reveal a future drug-disease relation. The paper is organized as follows: Section 2 revisits related work accompanied by our extensive investigation of embedded drug-disease associations in section 3. We close with conclusions in section 4.

2 Related Work

Research in the field of digital libraries has long been concerned with semantically meaningful similarities for entities and their relations. With a high degree of *manual*

curation numerous existing systems guarantee a reliable basis for value-adding services and research planning. On the one hand, automation can help to handle the *explosion* of scientific publications in this field, but on the other hand automation should not have a negative impact on quality, i.e. a high degree of precision has to be guaranteed. Arguably, the Comparative Toxogenomics Database (CTD[1]) is one of the best databases for curated relations between drugs and diseases. CTD contains both curated and derived drug-disease relationships. Because of the high quality, we use the curated relationships from CTD as ground-truth. Although manual curation achieves the highest quality it also comes with high expenses and tends to be incomplete [20]. In the past this led to the development of methods for automatic extraction of DDAs: ***Drug-centric****:* These approaches try to infer new and unknown properties (e.g. new application/side effect) of drugs from a drug-to-drug-similarity by means of chemical (sub-) structure (chemical similarity) [8]. ***Disease-Centric***: This approach calculates a similarity based on diseases and their characteristics. The hypothesis is: The same active substances can also be used for similar diseases (guilt-by-association rule, [14]). For example, phenotype information is compared to determine disease-similarity, whereby similar phenotypes indicate similar diseases. ***Drug Disease Mutual***: This approach is also known as the network-based approach and uses both, drug-centric and disease-centric approaches to derive/predict DDAs (see [4] for a good overview of different approaches). **Co-occurrence/mentioning**: Here, two entities are seen as similar and are thus related, if they co-occur within the same document. Moreover, co-occurrences in more documents of a collection speak for stronger entity relations [5]. The co-occurrence approach consists of two simple steps: 1) recognition of medical entities (through Named Entity Recognition) in documents (usually restricted to abstracts or even the same sentence) and 2) counting their common occurrences. Afterwards, counts can be used to infer DDAs. In our investigation, we also use the co-occurrence approach as a baseline for DDAs.

For the investigation of semantic relations between words, distributional semantic models are currently the state-of-the-art approaches [12]. The basic hypothesis is that words with a similar surrounding word context also have a similar meaning [18]. According to Baroni et al. [1] distributional semantic models (DSMs) can be divided into count-based models and predict models. Count-based models are generally characterized by (word) co-occurrence matrices being generated from text corpora. In contrast to count-based models, predict models try to predict the surrounding word-context of a word [1]. Compared to classical count-based models (e.g. LSA [16]), current, predict models such as Word2Vec presented by Mikolov et al.[13] lead to better results for predicting analogies as well as for other semantic tasks [1, 12]. Therefore, in our investigation we will rely on predict models as the state-of-the-art method for entity contextualization. In particular, we use the Word2Vec Skip-Gram model implementation from the open source Deep-Learning-for-Java[2] library.

With the increasing popularity of predict models, interest in the study of the semantic meaning of distance in high-dimensional spaces is growing. This is because of the non-

[1] http://ctdbase.org/

[2] https://deeplearning4j.org/

deterministic character of these models. State-of-the-art models such as Word2Vec use neural networks to predict contexts. To do this efficiently on large text corpora, random parameters are used, which however means that these methods are generally not deterministic. Therefore, it is rather difficult to decide whether a distance between entities always reflects a meaningful relation [15]. Elkes et al. [15] 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 [19] in order to determine for which distance a measured word pair reflects relatedness or associatedness. 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* has an effect on disease *y* or not [3]). In our investigation we measure the quality of this binary relation in the word embedding space.

3 Investigation of Embedded Drug-Disease Associations

We will first describe our pharmaceutical text corpus and basic experimental set-up decisions. Furthermore, we perform a ground-truth comparison and first show that DDAs are reflected in the embedding space. Here, we initially perform a comparison with the CDT data set followed by a text-mining co-occurrence comparison and we give the answers to our scientific questions Q1 and Q2 from Section 1. Then, we examine the predictive properties of the model with a retrospective analysis and show that DDA predictions are indeed possible and thus embedded DDAs can point to a future relation between drugs and diseases (Q3 and Q4).

Experimental Setup.

Evaluation corpus. PubMed[3] is with more than 27 million document citations the largest and most comprehensive digital library in the biomedical field. Since a full text access is not available for the most publications we used only abstracts for our evaluation corpus. With more training data, more accurate contexts can be learned. Thus, we decided to use a minimum of the 1000 most relevant abstracts for each entity (active substance). Whereby we relied on the relevance weighting of PubMed's search engine. Diseases as well as drugs often consist of several words (e.g. diabetes mellitus). This is a problem, because word embedding algorithms usually train on single words, resulting in one vector per word and not per entity. A solution to this problem is 1) recognize the entities in documents and 2) place a unique identifier at the entity's position in the text. For the recognition of the entities we used PubTator[4], a tool which is able to recognize pharmaceutical entities and returns a MeSH-Id for each of them.

Query Entities. As query entities for the evaluation, we randomly selected 350 drugs from the *DrugBank*[5] 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. As

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

[4] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/

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

ground truth, we selected for each drug all manually curated drug-disease associations from CTD. Moreover, we ensured that each selected drug has at least one manually curated drug-disease association in CTD.

Experiment implementation and parameter settings.

1. Text Preprocessing: Stop-word removal and stemming was performed using a *Lucene*[6] 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 preprocessing, word embeddings were created with DeepLearning4J's *Word2Vec*[7] implementation. To train the neural network, we set the word window size to 20, the layer size to 200 features per word and we used a minimum word frequency of 5 occurrences. Training iterations were set to 4. We tested several parameter settings but the above-mentioned turned out best.

3. 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 Experimental Investigation

First, we need to clarify how a relationship between drugs and diseases in embedding space can be qualitatively evaluated. We verify in the following tests that a DDA can be inferred on the basis of a linguistic distance and whether an unknown DDA is actually either an error or points to a future drug-disease association (DDA). In this context, the following quality criteria should be fulfilled:

- *Semantic Relationship Accuracy:* First, as the distance between a drug and a disease decreases, the likelihood for establishing a true DDA should increase. Furthermore, the distances of all true DDAs and the distances of all false ones should differ statistically significantly, so that the two sets are generally distinguishable. In addition, a sufficiently good quantitative result (i.e. high precision) should be achieved.
- *Prediction Accuracy:* Do embedded DDAs provide *more* or simply *different* information compared to text-mining co-occurrence approaches? Embedded DDAs should help to reveal a hidden context which is not expressed exclusively via co-occurrence in documents but via word-contexts. This hidden context should be meaningful which means *false* DDAs in time period *t* should become *true* DDAs in future publications (time period $t+E$).

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

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

3.2 Semantic Relationship Accuracy

In our first experiment we initially investigate whether a meaningful (true) DDA can be measured and in what degree of quality (precision). We also verify how a DDA-probability changes with a decreasing distance. For a measurement of DDAs, we first had to choose between a distance-threshold and a k-nearest-neighbors (k-NN) approach. Since distance-thresholds are difficult to determine in context-predicting models [15], we choose the k-NN approach. In our experiments we select the closest k-nearest disease neighbors (k-NDN) for each drug and measure the average precision for the following k's = 1, 3, 5, 10, 20.

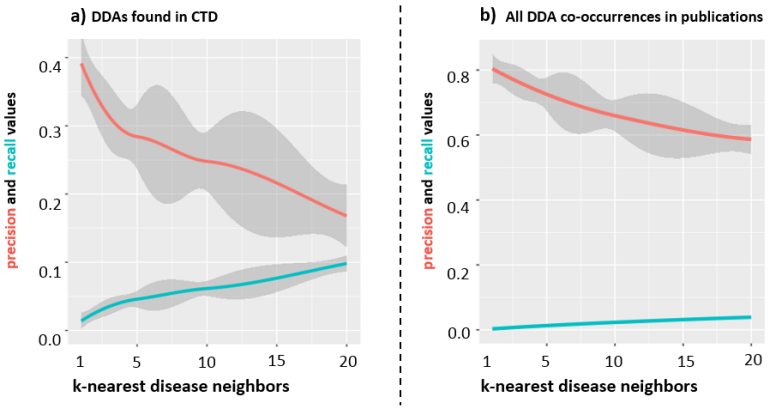

Fig. 1. Comparison of AVG-precision (red line) and AVG-recall (blue line) for different k-nearest disease neighbors using a) CTD and b) publications co-occurrence count and variance area in dark grey.

Figure 1 a) shows the results of our first experiment based on the CTD dataset. At k=1 the precision is 0.38 and drops to a value of 0.18 at k=20. The result confirms the hypothesis: With increasing distance the probability of a true DDA actually decreases. The CTD has a high quality due to the high manual effort but this is often accompanied by the disadvantage that a manually curated source is not complete and usually the most popular or most important DDAs are curated first [20]. Therefore we carry out an additional experiment (Figure 1b), with scientific publications as another comparison source which contains on the one hand (theoretically) all DDAs but is correct only under a co-occurrence assumption [5]: A DDA exists if an active substance *dr* and a disease *di* co-occur together within at least *x* publications. In addition, with a higher *x* the probability for a true DDA increases. For our experiments we set *x* on "at least 3" publications. Therefore, for a DDA embedding there must be at least 3 publications containing this pair. Only then do we count this co-occurrence as a true DDA. We repeated our first experiment with the new source and present the results for the k-nearest disease-neighbors of each drug in figure 1b. At k=1 the precision is at ~0.80 and drops to ~ 0.60 at k=20. We achieve an AVG precision of ~0.7 for the different k's. Like in the previous experiment this experiment supports the hypothesis that the probability for a

true DDA decreases with an increasing distance. And secondly, we can assume accurate precision-values within the first 20 k-nearest disease neighbors (k-NDN).

With our next experiment we would like to answer the following questions: Which AVG cosine-similarity values have true DDAs compared to false DDAs? Is there a general and meaningful threshold that can help us to distinguish between true and false DDAs? Are there statistically significant differences in these quantities? To test this we first measure all diseases for each of the 350 drugs up to a minimum cosine-similarity of 0.001. In total, we obtain 184,553 DDAs in this way. To analyze the DDA pairs and to answer the questions we use a histogram, density, boxplot, and a Welch two sample t-test.

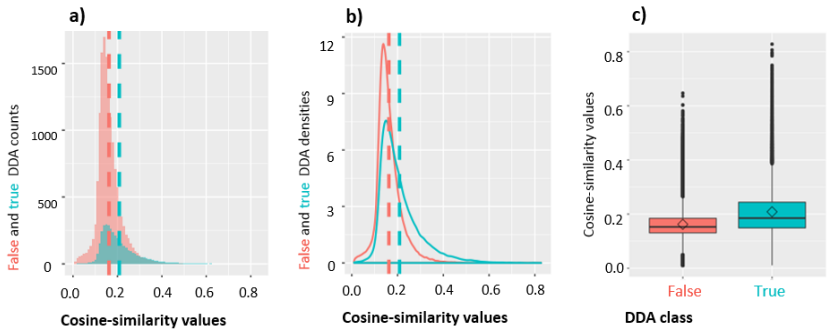

Fig. 2. Histogram, density and boxplot for all nearest-disease neighbors. False DDAs (red) and true DDAs (blue) as well as their means (dotted lines).

Can we find meaningful similarity-based thresholds? As can be seen in Figures 2a, 2b, and 2c (diamond represents the means), the amounts of true (blue) and false (red) DDAs overlap strongly. The majority of all true DDAs share a large amount of similar similarity values with the set of false DDAs. The result suggests the assumption that (1) the DDAs are generally difficult to separate by a *cut-off* value and (2) in general DDAs are surprisingly poorly represented by the word embedding model. *Are the two sets significantly different?* To prove this we performed a Welch two sample t-test with a confidence interval of 99%. With a p-value < 2.2e-16 we are significantly below the threshold of 0.05 and thus there is a significant difference between the two sets.

Our previous experiment suggests that both sets of true and false DDAs cannot be meaningfully separated (e.g. many false positives) with a similarity threshold, even though the sets differ significantly. On the other hand, our co-occurrence experiment (see Fig. 1b) promises an adequate precision (between 60% and 80%) with a maximum of 20 disease neighbors. For this purpose we limit the k-NDN in our next experiment again and investigate whether the sets can be better separated using a smaller k. The figures 3a and 3b show the results for the k=20 next disease neighbors. Compared to the results shown in figure 2, the figures in 3 exhibit that the distributions of true and false DDAs can be separated better. We have a larger proportion of true DDAs (~60%) compared with the portion of true DDAs from our second experiment (~0.35, using density curve intersection cut-off point). Experiment 1-3 indicate that a (adequate) meaningful result can only be achieved with smaller *k's*.

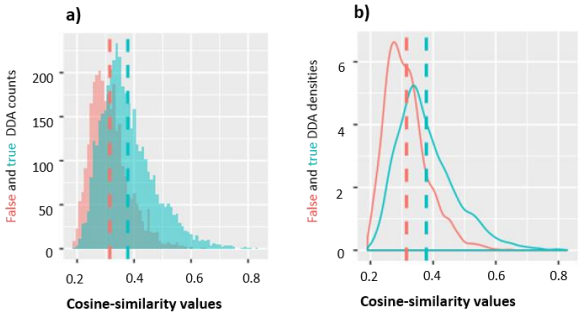

Fig. 3. Histogram and densities for only 20 nearest disease neighbors. False DDAs (red) and true DDAs (blue) as well as their means (dashed lines).

3.3 Prediction Accuracy

In the first experiment section we proved that distance actually correlates with correctness for at least 20 nearest disease neighbors. What gives us hope for possible predictions of DDAs? Word2Vec is not based on co-occurrence but on learning contexts from millions of publications. The consequence of this learning is that entities with a similar context are probably closer together in vector space. This property can be used and transferred on DDAs. Which means: A similar context points to an intrinsic relation between drugs and diseases. *Can new relationships (DDAs) be predicted with this property and how can we measure it?* A possible approach to verify this is a retrospective analysis and the determination of the proportion of all false DDAs at time *t* (don't co-occur in publications) that become *true* at time *t+E* (co-occur in publications). Here, we refer to this type of entities as *future* DDAs. This experiment requires adjustments to the evaluation corpus and to the evaluation implementation:

Evaluation corpus: In order to calculate the change in different time periods we divide our previous corpus into four corpuses: 1900-1987, 1900-1997, 1900-2007, 1900-2017. Each corpus contains only the documents for the respective time period.

Evaluation implementation: Now we train our model with each time period using the same parameters as described in the previous experiments. Afterwards, we first check the proportion of DDAs that are true in time period *t*. Then we check how the precision changes when we measure in time *t+E* (next time period). *How fast grows the prediction of DDAs compared to the general increase of DDAs found in literature?* To enable a comparison of increase we generate the two control groups *R2* and *R3*. For *R3*, we identify all *future* DDAs in time period *t* and replace the disease with a random selected disease. Then we measure the proportion of *true* DDAs at *t+E*. With *R3* we investigate how big the difference to a randomly *coming true* of a DDA really is. For control group *R2* we replace the future DDAs with randomly selected disease neighbors (at a range between 21- and 40-NDNs) and compare the difference to *R1* (20-NDSs). With R2 we want to investigate if with increasing distance the probability for a DDA-prediction decreases. We calculate AVG precision for the following k=1, 3, 5, 10, 20 again. Using this approach we compare all time periods:

Table 1. Avg. precision values for the different time periods. Results of 20-NDNs (*R1*), (*R2*) results of control group with replaced future DDAs with nearest disease neighbors (range 21 to 40-NDNs), (*R3*) results of control group with replaced future DDAs with randomly selected diseases. Best predictive results in bold.

Time periods	1900-1987	1900-1997	1900-2007	1900-2017
1900-1987	R1: 0.566 R2: 0.566 R3: 0.566	**R1: 0.573** R2: 0.568 R3: 0.566	**R1: 0.589** R2: 0.573 R3: 0.567	**R1: 0.661** R2: 0.610 R3: 0.570
1900-1997	x	R1: 0.623 R2: 0.623 R3: 0.623	**R1: 0.628** R2: 0.626 R3: 0.623	**R1: 0.680** R2: 0.644 R3: 0.625
1900-2007	x	x	R1: 0.671 R2: 0.671 R3: 0.671	**R1: 0.685** R2: 0.677 R3: 0.671
1900-2017	x	x	x	R1: 0.700 R2: 0.700 R3: 0.700

Table 1 shows the results of our retrospective analysis. As we can see, AVG precision increases measurably for all subsequent time periods. In fact, each time period actually contains DDAs that will appear in future publications. The largest percentage increase of future DDAs can be observed for the time period 1900-1987 (increased from 0.566 to 0.661). Whereby in this case future DDAs denote DDAs contained in the period 1900-2017 (contains all known/true DDAs) minus all DDAs contained in the period 1900-1987. Taking these future DDAs into account, the AVG precision increases for the time period 1900-1987 by 17% from 0.566 to 0.661. Thus, the best results are achieved for the longest time period (30 years). Compared to control group *R3* (random disease swap) there is a remarkable difference (total increase till 2017 is less than 1%) between a randomly coming true of a future DDA. R2 (swap with 21-40 NDNs) has a slower prediction increase compared to *R1*, which means: With increasing distance the probability for a DDA prediction decreases. It can be seen that for *R1* there is always growth and therefore predictions can always be made using the context-based method. In short, the procedure is able to predict DDAs. Surprisingly, there is also a pattern in the columns: The column values in column 1900-2017 hardly differ. For example the standard deviation in the column is only ~ 0.016. *Is there a certain similarity-distance area where future DDAs concentrate?* After showing that DDA predictions are possible in general, our next step is to investigate the distance range of DDAs that are currently false but will become true in future time periods. Knowing in which area these future DDAs are located can help us to find more useful cut-offs for later applications (e.g. DDA prediction). For this purpose we analyze the position of median, mean, and also the distribution density of the 20 nearest disease neighbors in detail. As time periods we choose 1900-1987, 1900-1997, 1900-2007, and analyze the proportion of *false* DDAs that will become *true* by the end of 2017. Figure 4 shows the results of our last experiment. Row 1 shows the results of the distribution and row 2 the density of true

DDAs (green), false DDAs (red), and future true DDAs (blue). Each column in figure 4 represents the results of the different time periods 1900-1987, 1900-1997, and 1900-2007. In all figures, the mean values are shown as dotted lines. The median values for the different periods and DDA types are listed in table 2.

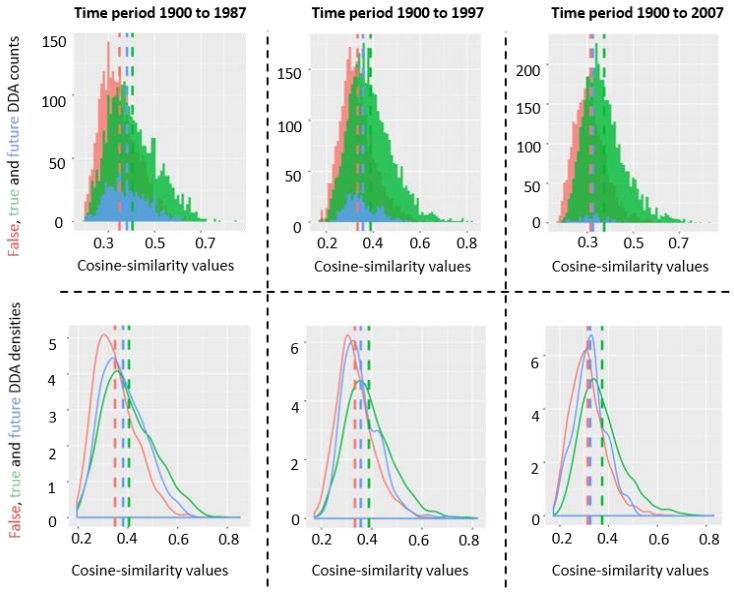

Fig. 4. False DDAs (red), true DDAs (green) and future DDAs that will co-occur in publications by the end of 2017 (blue). Dashed lines represent the mean values of the respective sets.

As can be seen in the distributions of Figure 4, the proportion of future- as well as false DDAs decreases over time and the proportion of true DDAs increases. Surprisingly, both mean (Figure 4) and median (Table 2) values remain relatively stable, although for example the corpus of time period 1900-1987 is more than twice as large as the corpus of 1900-2007 and the proportion of true and false DDAs changes remarkably.

Table 2. Median values for different time periods and DDA types.

Time period	True DDA	Future DDA	False DDA
1900-1987	0.39	0.37	0.34
1900-1997	0.38	0.34	0.32
1900-2007	0.36	0.33	0.31

In addition, mean and median values of the future DDAs always lie between the false and the true DDAs. Furthermore, there is always a stable ~5% distance between false and true DDAs. Thus, the corpus size as well as the proportion of true and false DDAs seem to have less effect on the distance between these two groups.

4 Conclusion and Future Work

We investigated in this paper if relations between embedded pharmaceutical entities are reflected in high-dimensional space and if their distance to each other correlates with the probability of their (binary) relationship. This is currently an important research question because non deterministic Word-Embedding models, like Word2Vec are on the one hand popular models for semantic as well as analogy tasks but on the other hand their properties are not fully investigated. Questions like the following remain: What has a model actually learned? How can we assess whether a result is meaningful? Answering these questions is essential, before we can use context-predicting models for scientific-entity relations like DDAs in innovative digital library services.

In this context we first proved that with an increasing distance the probability for a DDA decreases. We have shown that a sufficient AVG-precision can be achieved with 20-nearest-disease neighbors (NDNs) of an active substance. For example, the AVG-precision ranges between 60% for the 20-NDNs and 80% for the first one. Therefore, the threshold of 20-NDNs might be a good choice for a selection of DDAs because it is a trade-off between sufficiently good precision values and the number of future DDAs included in this set. Surprisingly most DDAs can't be distinguished with a distance-threshold because most false and true DDAs have a similar distance and the two sets overlap. We concentrated our investigation on the 20-NDNs and demonstrated that sets of false and true DDAs can be separated better. That, since many false DDAs have similar distances as true DDAs, led us to the question: Is it probable that a hidden and meaningful DDA relation has been learned? We investigated this question with a retrospective analysis. We have shown that a significant proportion of the false DDAs will actually become true in the future (increase of up to ~ 17%). Additionally we have shown that with decreasing distance the probability for a DDA prediction increases. These results open up the possibility of predicting DDAs on the basis of co-contexts found in literature, which in turn can be of important benefit in the field of drug repurposing. Afterwards, we examined the distance range of this specific entities which we call *future* DDAs. We have shown that restricted to the 20-NDNs mean and median of future DDAs lies always in between the mean and median of false and true DDAs. The positions of the various values determined (median, mean) remain relatively stable over the periods and for different corpus sizes. In our future work we will examine if embedded diseases can reveal the intrinsic relationship between groups of embedded drugs investigated in our previous work [7]. And secondly, we want to explore the possibilities of DDA predictions and compare this method with state-of-the-art text mining approaches for DDA predictions.

References

1. Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 238-247).

2. Leaman, R., Islamaj Doğan, R., & Lu, Z. (2013). DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, *29*(22), 2909-2917.
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. Lotfi Shahreza, M., Ghadiri, N., Mousavi, S. R., Varshosaz, J., & Green, J. R. (2017). A review of network-based approaches to drug repositioning. *Briefings in bioinformatics*, bbx017.
5. Jensen, L. J., Saric, J., & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews genetics*, *7*(2), 119.
6. Dudley, J. T., Deshpande, T., & Butte, A. J. (2011). Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics*, *12*(4), 303-311.
7. Wawrzinek, J., & Balke, W. T. (2017). Semantic Facettation in Pharmaceutical Collections Using Deep Learning for Active Substance Contextualization. In *International Conference on Asian Digital Libraries* (pp. 41-53). Springer, Cham.
8. Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., & Whaley, R. (2009). Predicting new molecular targets for known drugs. *Nature*, *462*(7270), 175.
9. Agarwal, P., & Searls, D. B. (2009). Can literature analysis identify innovation drivers in drug discovery? *Nature Reviews Drug Discovery*, *8*(11), 865.
10. Ngo, D. L., Yamamoto, N., Tran, V. A., Nguyen, N. G., Phan, D., Lumbanraja, F. R., & Satou, K. (2016). Application of word embedding to drug repositioning. *Journal of Biomedical Science and Engineering*, *9*(01), 7.
11. Lengerich, B. J., Maas, A. L., & Potts, C. (2017). Retrofitting Distributional Embeddings to Knowledge Graphs with Functional Relations. *arXiv preprint arXiv:1708.00112*.
12. Mikolov, T., Yih, W. T., & Zweig, G. (2013). 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* (pp. 746-751).
13. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
14. Chiang, A. P., & Butte, A. J. (2009). Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clinical Pharmacology & Therapeutics*, *86*(5), 507-510.
15. 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.
16. Dumais, S.T. (2004). Latent Semantic Analysis. In *Annual review of information science and technology (ARIST)*, Vol. 38(1), Association for Information Science & Technology.
17. Larsen, P. O., & Von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, *84*(3), 575-603.
18. Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes*, *6*(1), 1-28.
19. Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.
20. Rinaldi, F., Clematide, S., & Hafner, S. (2012, April). Ranking of CTD articles and interactions using the OntoGene pipeline. In *Proceedings of the 2012 BioCreative Workshop*.