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

Janus Wawrzinek
Institute for Information Systems
TU-Braunschweig
Braunschweig, Germany
wawrzinek@ifis.cs.tu-bs.de

José María González Pinto
Institute for Information Systems
TU-Braunschweig
Braunschweig, Germany
pinto@ifis.cs.tu-bs.de

Wolf-Tilo Balke
Institute for Information Systems
TU-Braunschweig
Braunschweig, Germany
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 tool for this problem is to better understand important pharmaceutical entities and their relations to each other in order to facilitate access to the embedded knowledge in digital libraries. State-of-the-art approaches in the field of neural-language models (NLMs) enable progress in learning and the prediction of such relations in terms of semantic quality, scalability, and performance. However, the advantages go hand in hand with disadvantages regarding the interpretation of a neural network outcome, e.g.: *When is a learned entity relation meaningful?* In this context NLMs are hard to convey in an academic environment and for academic searches, when even such an elementary question remains unclear. Based on a pharmaceutical use case, we investigate this question for important pharmaceutical entity relations in the form of drug-disease associations (DDAs). In our experiments we first show that unusual hyperparameters can lead to an improvement in semantic quality of up to 51%, both in learning existing associations and in predicting new ones. Afterwards, we show that there exist strong correlations between corpus characteristics and the number of embedded entity relations to be extracted from the embedded space. This allows us to determine entity-specific thresholds that are independent of a similarity value, which is usually difficult to estimate and to interpret. Our entity-specific thresholds open up the possibility towards the construction of (meaningful) embedded pharmaceutical networks, which can be used to explore and to explain the learned relationships between pharmaceutical entities.

CCS CONCEPTS

• **Information systems~Digital libraries and archives**
• **Information systems~Specialized information retrieval**

KEYWORDS

Neural language models, semantic similarity, similarity thresholds.

1 INTRODUCTION

Apart from being confronted with massive document holdings, today's medical digital libraries also have to manage their exponential growth [5]. For example, the search for the term "*Diabetes*" in the PubMed medical digital library leads to a result of over 40,000 documents for the year 2018 only. Due to this problem of increasing unmanageability, we need efficient methods and new access paths to make digital libraries more usable and useful in such research-related areas [2]. One of the most effective ways to address this problem is to automatically extract important pharmaceutical entities such as active ingredients, diseases, and their relationships to reveal the embedded knowledge contained in such massive collections.

Numerous works have recognized this trend and focus on the automatic detection of these entities and their relationships known as Drug-Disease Associations (DDAs) [7, 8, 10]. *What is a DDA?* A DDA exists when a) a drug helps against a certain disease (cures, prevents, alleviates) or b) a drug induces a disease in the sense of a side effect [8, 16]. *Why are DDAs of interest?* For instance, DDAs are considered possible candidates for drug-repurposing [3]. Here, pharmaceutical research pursues the interest to use well-known and proven active substances against other diseases, as this generally leads to a lower risk in terms of negative side effects.

Based on the interests mentioned above, numerous computer-based methods were developed to derive DDAs from text corpora as well as from specialized databases [14]. The similarity between active substances and diseases forms the basis here, and numerous popular methods exist for calculating a similarity between these pharmaceutical entities such as chemical (sub) structure similarity [15] or network-based similarity [16]. Scientific literature is one of the primary sources in the investigation of new drugs [10], which is why newer approaches use Distributed Semantic Models (DSMs) to calculate linguistic or lexical similarities between entities in order to deduce their properties and relationships [1, 9].

The use of DSMs in this area is based on the following (context) hypothesis [1]: If entity *A* appears in the word context (sentence, abstract) with entity *B* and if entity *B* appears in the word context with entity *C*, then this transitive (hidden) common context, between *A* and *C* via entity *B*, is recognized. Usually, these entities are embedded in a high dimensional space (e.g. 100-400

dimensions) and spatially positioned closer to each other due to numerous similar common contexts [1]. In our previous work [17] we investigated this hypothesis and surveyed the embedded DDAs. Not only were we able to show the extent to which distance correlates with the probability of a DDA, but we were also able to confirm the hypothesis of predictability by means of a retrospective analysis. *Can the number of meaningful DDAs and predictions be increased?* Hyperparameters and training corpora are usually the first starting points to answer this question. The study of the relationship between hyperparameters, corpus and the semantic quality of these models [19, 20, 24, 29, 30] grows with the increasing popularity of state-of-the-art DSMs such as Word2Vec [11]. As this has been insufficiently investigated for important scientific entities and their relations such as the Drug-Disease Associations, we close this gap with our work and focus not only on semantic quality, but also on the quality of DDA-predictions as well as on entity-specific thresholds.

Contribution: Our investigation is divided into two main sections. In the first section we examine the effect of hyperparameters on the semantic quality of embedded DDAs. The research questions we investigate and answer here are as follows: (***Q1***) *Do important hyperparameters such as window size and dimension size have a substantial qualitative influence on the detection and prediction of DDAs in embedded space? (**Q2**) What influence do the hyperparameters have for different time periods and corpus sizes?*

Our results show that the choice of standard parameters, especially window size, for DDAs leads to the worst semantic quality results and predictions. We show that unusual window sizes can lead to an increase in semantic quality of more than 51%. We can extract more meaningful DDAs from the embedding space and by retrospective analysis we can also show that on average we can predict more new DDAs. An additional finding is that with increasing data (corpus size) and window size, semantic quality increases substantially.

These results lead us to our second section and the following hypothesis: *As the amount of data per entity increases, the semantic quality also increases. Therefore, as the amount of entity related data increases, more meaningful DDAs can be extracted from the embedding space for a certain entity.* To confirm the hypothesis, we investigate the following research questions: ***(Q3)*** *Does a relationship exist between corpus data volume and semantic quality per embedded entity that would indicate a correlation? **(Q4)** Is this correlation stable in different time periods and for different corpus sizes? **(Q5)** Can we determine entity specific thresholds based on this correlation? **(Q6)** If so, are these entity-specific thresholds more effective than conventional methods such as k-NNs or Similarity Thresholds?*

In this paper we first show that the number of DDAs contained in the documents correlates strongly with the probability of a meaningful DDA in the embedded space. In addition, this correlation is stable over different time periods and corpus sizes (standard deviation 0.018). However, surprisingly this *stability property* only applies in combination with high window sizes. Furthermore, we show that we can use this correlation to train a classifier for predicting entity-specific thresholds. This enables us to become independent from the difficult to determine k-NN or Similarity-Thresholds [17, 34]. Hereafter, we show that with our entity-specific thresholds we can extract a multiple (up to 142%) of meaningful DDAs from the embedded space as compared to the conventional approaches (k-NN, Similarity-Threshold). The resulting entity-specific thresholds allow us to build more meaningful links between embedded drugs and diseases, but in addition they enable us to develop a concept for a contextualized drug-disease network (Section 4).

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. We also believe that our method can be generalized to other entity classes (e.g. drug-targets, genes) and can help to improve the semantic quality and prediction of entity associations. This can be used to generally improve approaches that use word embeddings for query expansion [31], in information retrieval tasks [32] as well as approaches that use neural embeddings as features in deep learning models [33]. The paper is organized as follows: Section 2 revisits related work accompanied by our extensive investigation of embedded drug-disease associations in section 3. In section 4 we outline the idea of a Contextualized Pharmaceutical Network. We close with conclusions and future work in section 5.

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. 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 [25]. 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, [3]). 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

[1] http://ctdbase.org/

both, drug-centric and disease-centric approaches to derive/predict DDAs (see [16] 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 suggest stronger entity relations [26]. 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 [6, 11]. The basic hypothesis is that words with a similar surrounding word context also have a similar meaning. According to Baroni et al. [6] distributional semantic models (DSMs) can be divided into count-based models and predict models, which are also known as Neural Language Models (NLMs). Count-based models are generally characterized by (word) co-occurrence matrices being generated from text corpora. In contrast to count-based models, NLMs try to predict the surrounding word-context of a word [6]. Compared to classical count-based models (e.g. LSA [4]), current NLMs such as Word2Vec presented by Mikolov et al. [11] lead to better results for predicting analogies as well as for semantic tasks [1, 35]. 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 CBOW performs better with syntactic tasks (large-bigger), while the Skip-gram model is better suited for semantic tasks [11]. 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-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 [24]. Elekes et al. [24] 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 has an effect on disease y or not [8]). In our investigation we measure the quality of this binary relation in the word embedding space.

Not only hyperparameters and the corpus have an effect on the similarity values between entities, but also the question whether two entities are similar, in the sense of being synonymous, or (rather) associated. The linguistic community has been dealing with this question for quite some time. This led to the discussion about, and disambiguation of, *Concept-Similarity* and *Concept-Associatedness* [19]. According to Hill et al. [19], for example, the concept pair (*car, bike*) is rather similar, while for the pair (*car, fuel*) it is rather an associative relationship. Determining this difference is difficult, that is why usually hundreds of curators have to manually evaluate how similar or associated different pairs are. In this context, we investigate entity relations that are rather (exclusively) associated, because as mentioned above, drugs can be similar to each other, e.g. through a similar chemical structure [8] or a similar therapeutic effect. Diseases can also be similar to each other, e.g. due to similar phenotypes or genetic similarity [18]. On the other hand, the combination of active substances and diseases is considered a drug-disease association [8].

Once the question of "*which pairs of objects*" can be used to carry out a DSM quality evaluation has been clarified, the next question is which evaluation approach should be used. In order to measure the quality of word embeddings, *intrinsic* and *extrinsic* approaches can be distinguished [19]. Intrinsic approaches are mostly based on the comparison of DSM word vectors (and their nearest neighbors) with manually curated word pair lists (a good overview can be found here [19]). These approaches examine which DSM models and their hyperparameters (e.g. window size, dimension size) generally lead to good performance values in comparison, whereby 200-400 dimensions and a window size of 5 are used in general [6]. In contrast to intrinsic approaches, extrinsic approaches try to determine the best hyperparameters for a given semantic task. The fact that extrinsic approaches should be preferred is because there may be qualitative differences per DSM task [20] and we cannot always infer extrinsic quality from an intrinsic evaluation [21]. Thus we perform an extrinsic evaluation and our semantic task is the recognition of existing and the prediction of new entity associations. Once the best parameters have been found, the next question is whether meaningful similarity thresholds exist for semantically similar or associated pairs. However, similarity thresholds are difficult to calculate and can vary for different models [22, 24] due to dependence on hyperparameters and corpus. In [17] we have shown that even similarity thresholds calculated per model are problematic because the sets of true DDAs and false DDAs overlap strong. 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.

[2] https://deeplearning4j.org/

3 EXPERIMENTAL INVESTIGATION

We will first present the used terminology. Afterwards we describe our pharmaceutical text corpus and basic experimental set-up decisions. The sections (3.1 and 3.2) base strongly on our previous work [17]. Our investigation is divided into two sections. In the first section we first investigate the effects of hyperparameters (dimension size and window size) and answer the research questions Q1 and Q2 formulated in the introduction. To investigate *Q1-Q2* we use the manually curated DDAs from CTD as well as the described co-occurrence approach. The co-occurrence approach also allows us to perform a retrospective analysis over a time-period of 118 years.

In the second section, we focus on the relationship between corpus data volume and semantic quality per embedded entity as well as we investigate entity specific thresholds. In this context we give answers for the questions *Q3-Q6*.

As shown in [17] on the one hand DDA predictions are possible but on the other hand the majority of true DDAs can't be extracted from the embedding space. Therefore, we focus in our investigation on precision as the main quality-measure.

3.1 Terminology

In our experiments we train different Word2Vec models and extract a set of k-Next-Disease-Neighbors (hereafter **k-NDNs**) per drug entity from the resulting embedded spaces. A particular drug and one element of the k-NDN set each form a Drug Disease Association (**DDA**). DDAs are ***true* DDAs** if they can also be found in the ground truth set, otherwise they are ***false* DDAs**. Since we perform a retrospective analysis, a *false* DDA may turn into a *true* DDA, if it can be found in a "future" ground truth set. These DDAs are referred to as ***future* DDAs**. *True* DDAs together with *future* DDAs are also referred to as *meaningful* DDAs.

3.2 Experimental Set-Up

Evaluation corpus. *PubMed*[3] is with more than 28 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. Furthermore, we collected abstracts for the time period between 1900-01-01 and 2018-01-01. 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 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.

3.3 Experimental Implementation and Parameter Settings

In the following we describe experimental implementation details and parameter settings.

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.

Word Embeddings. After pre-processing, word embeddings were created with DeepLearning4J's *Word2Vec*[7] implementation. To train the neural network, we set the word window size in our investigations between 5 and 50, the layer size between 100 to 400 features per word and we used a minimum word frequency of 5 occurrences.

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).

3.4 Experimental Investigation

First, to answer *Q1* and *Q2* we need to clarify how an improvement of a relationship between drugs and diseases in embedding space can be qualitatively evaluated. Thus, in the first experiments we will investigate the effect of hyperparameters (dimension size/window size) on embedded DDAs. In this context the following quality criteria have to be fulfilled:

- *Semantic Relationship Accuracy Improvement*: If hyperparameters have an influence on the semantic quality of embedded DDAs, then this should lead to measurable qualitative differences for different parameter settings. This can be expressed, for example, in substantially increased AVG precision values (e.g. more true-DDAs with the same number of k-NDNs).
- *Semantic Relationship Accuracy Stability:* Do certain parameter settings always lead to better results (e.g. more meaningful DDAs) in different time periods and for different corpora? The influence of the parameters should remain stable. Therefore, a parameter setting S_1, which leads to

[3] https://www.ncbi.nlm.nih.gov/pubmed/
[4] https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/
[5] https://www.drugbank.ca/
[6] https://lucene.apache.org/
[7] https://deeplearning4j.org/word2vec

better results than a parameter setting S_2, should always lead to better results in every time period and all corpora.

Based on the previous findings, we investigate for *Q3* whether there is a general relation between corpus data and the number of meaningful embedded DDAs. For this, the following quality criterion should be fulfilled:

- *Corpus-Data and Embedded-Entity Relationship*: If there exists a relationship between corpus data volume and the embedded drug entities, then the probability of extracting more meaningful DDAs should increase as the entity-data volume increases. This can be expressed, for example, by the fact that the AVG precision with *k*-NDNs increases with increasing entity-data volume. This relationship should also hold for all investigated time periods and corpora.

Afterwards, we investigate the correlation for Q4-Q6 in order to derive entity-specific thresholds. In this context the following quality criteria should be fulfilled:

- *Correlation and Correlation Stability*: There is a chance to derive possible thresholds from a correlation. However, a strong correlation (Pearson coefficient > 0.7) between corpus data set and embedded entities is preferable for this task. At best, this correlation also exists in different corpora and time periods and shows a small standard deviation.
- *Threshold Accuracy*: If thresholds can be determined on the basis of a corpus-data correlation, then these should lead to better results compared to standard approaches (Similarity-Threshold, fixed-k-NDNs). Precision as a measure of comparison alone is less meaningful because it can be easily increased by setting similarity values high or by extracting fewer k-NDNs. From a user perspective, it should provide a *moderate* AVG precision. Therefore, for a given precision, the specific threshold should lead to higher recall than the comparative approaches with the same precision allow.

3.5 Semantic Relationship Accuracy Improvement

In our first experiment we initially investigate whether the amount of true DDAs can be increased using different hyperparameters and in what degree. For a measurement of DDAs, we choose a k-nearest-neighbors (k-NNs) approach. In our first experiments we select the closest k-nearest disease neighbors (k-NDNs) for each drug and measure the average precision and recall for different k's (where $k > 0$ and $k <= 20$). We perform the experiment for changing window sizes (w=5, 10, 20, 50) and afterwards for different dimension sizes (d=100, 200, 300, 400). In order to evaluate the various window sizes, we have set the dimension size in the tests to 200 dimensions [19] as recommended for Word2Vec.

Figure 1a) shows the results of our first experiment based on the CTD dataset. For example at k=1 and a window size w=20 the AVG precision is 0.38 and drops to a value of 0.18 at k=20. We can see that with increasing window sizes the probability of a true DDA also increases. The increase in AVG precision between the general default setting (w=5) and our maximum window size (w=50) is 51% and AVG recall increases by 45%. Thus, we can extract substantially more true DDAs from the embedded space with an increased and rather unusual window size.

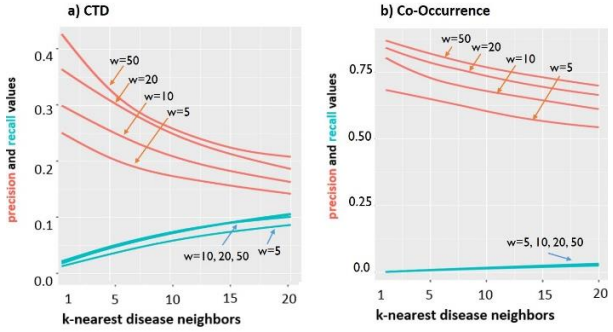

Figure 1. Comparison of AVG-precision (red line) and AVG-recall (blue line) for different k-nearest disease neighbors using a) CTD and b) Co-Occurrence count.

The CTD is of 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 [25]. 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 [26]: A DDA exists if an active substance *dr* and a disease *di* co-occur 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. The results for the k-NDNs of each drug are presented in Figure 1b. Here, for example at k=1 and a window size w=50 the AVG precision is 0.88 and drops to a value of 0.71 at k=20. The increase in AVG precision between the general default setting (w=5) and our maximum window size (w=50) is 27% and AVG recall increases by 29%. Thus, we can extract again substantially more true DDAs from the embedded space with an increased and rather unusual window size.

Result interpretation. Hill et al. [19] have investigated the hypothesis that larger window sizes may be better at learning associations, as mentioned in [27]. However, their results show that with an increased window size (from 2 to 10) the results for associations become even (negligibly) worse. Also Elekes et al. [24] have shown in their experiments a small influence of window size on similarity values in their models. In [21] it was shown that with an increasing window size and an intrinsic evaluation the performance increases, whereas in an extrinsic evaluation the performance decreases with an increasing window size. The work and the controversial results confirm the assumption that the best parameters must be found for each semantic task [19]. In this

controversial discussion we can identify the following laws in our experiments and our semantic task: We have not only shown that with increasing window size the values become substantially better, but that *unusual* window sizes lead to this improvement. Larger window sizes seem to learn the abstract associations between drugs and diseases much better. Our conclusion for pharmaceutical entities is: 1) It is worthwhile to test window sizes beyond the commonly used 2-10 (as tested in [19, 27, 28]). 2) *Forget the dots*: For larger window sizes it is worth breaking the boundaries of a sentence. In order to train on larger window sizes, we have used the entire abstract instead of only the sentences.

Next, in the following experiment we investigate the effect of dimension size on the semantic quality of DDAs. We repeat the evaluation with CTD and the co-occurrence approach. As window size we use the default setting of w=5 in this experiment with a different dimension size d (where d=100, 200, 300, 400).

Table 1. Comparison of AVG-precision and AVG-recall using CTD.

Dimension size	Precision	Recall
100	0.16	0.030
200	0.18	0.033
300	0.19	0.034
400	0.18	0.033

Our results are presented in Table 1 for CTD and in Table 2 for co-occurrence count. As can be seen in Table 1 and 2, the changed dimension size has hardly any influence on the semantic quality. For co-occurrence count a setting of 200 dimensions seems to already lead to the best values. In contrast to the window size, our dimension size result is similar to works like [24] and [29], where it was found that the models are quite similar with different dimension settings.

Table 2. Comparison of AVG-precision and AVG-recall using publications co-occurrence count.

Dimension size	Precision	Recall
100	0.58	0.013
200	0.61	0.014
300	0.60	0.014
400	0.60	0.013

3.6 Semantic Relationship Accuracy Stability

In this section, we examine how different parameter settings affect DDA quality for different time periods and corpus sizes. However, based on the dimension size results, we will focus exclusively on the effect of window size on accuracy stability properties and the prediction of DDAs. In [17] we not only have shown that DDA predictions are possible using NLMs, but we have also compared NLM results with two control groups. Among other things we have shown that there is a clear difference to a random guess of a DDA. In this context we do not repeat our experiments with the control groups from [17], because they would not lead to new findings.

*(**Q2**) What influence do hyperparameters have for different time periods and corpus sizes?* A possible approach to measure this is a retrospective analysis [17] 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 corpora: 1900-1988, 1900-1998, 1900-2008, 1900-2018. Each corpus contains only the documents for the respective time period.

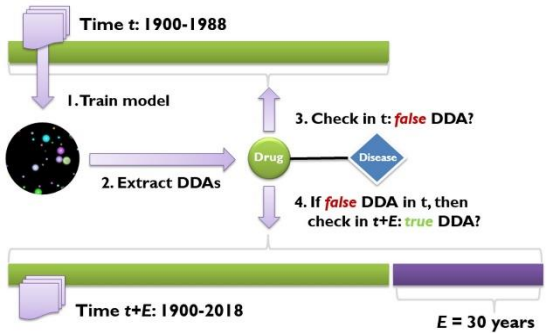

Figure 2. Retrospective Analysis approach overview.

Evaluation implementation: Now we train our model with each time period (Fig. 2, step 1) using different window size parameters. As next, we extract DDAs from the embedded space (Fig. 2, step 2). Afterwards, we first check the proportion of DDAs that are true in time period t (Fig. 2, step 3). Then we check how the precision changes when we measure in time t+E (next time period) and this for different window size settings (Fig. 2, step 4). We calculate AVG precision for different k's (where $k > 0$ and k<=20) again and use the following window sizes w= 5, 10, 20, 50. Using this approach we compare all time periods with different window size settings. Table 3 shows the results of our retrospective analysis with varying window sizes.

As we can see (Table 3), AVG precision increases measurably for all subsequent time periods as well as with increasing window size. With increasing window size we can also assume a higher future precision. The larger the corpus (diagonal), the higher the precision, the more meaningful DDAs we can extract on average. The larger the corpus, the higher the AVG-precision difference between the different window sizes. In the following section we investigate the relationship between corpus size and DDA precision further.

Table 3. AVG precision values for the different time periods. Best (predictive) AVG precision results in bold.

Time periods	1900-1988	1900-1998	1900-2008	1900-2018
1900-1988	W5: 0.489 W10: 0.526 W20: 0.560 **W50: 0.596**	W5: 0.497 W10: 0.533 W20: 0.566 **W50: 0.604**	W5: 0.517 W10: 0.552 W20: 0.585 **W50: 0.625**	W5: 0.601 W10: 0.638 W20: 0.667 **W50: 0.700**
1900-1998	x	W5: 0.522 W10: 0.588 W20: 0.629 **W50: 0.672**	W5: 0.535 W10: 0.597 W20: 0.638 **W50: 0.680**	W5: 0.586 W10: 0.651 W20: 0.693 **W50: 0.735**
1900-2008	x	x	W5: 0.589 W10: 0.653 W20: 0.707 **W50: 0.746**	W5: 0.610 W10: 0.671 W20: 0.722 **W50: 0.760**
1900-2018	x	x	x	W5: 0.605 W10: 0.685 W20: 0.737 **W50: 0.771**

3.7 Corpus-Data and Embedded-Entity Relationship

So far, we have performed the experiments with a maximum of 20 NDNs. Here, we can observe that the AVG precision varies considerably for different active substances. The standard deviation for 20-NDNs is 23%. Therefore, for some active substances the choice of more than the 20-NDNs seems appropriate, while for others fewer NDNs would probably make more sense. *Can we estimate the number of NDNs per drug entity?* It seems reasonable to determine a similarity-value threshold.

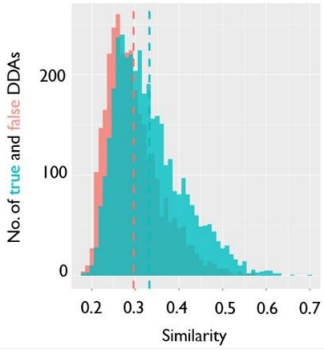

Figure 3. Distribution of true (blue) and false (red) DDAs for 20-NDNs per drug-entity. Trained with *w*=5 and *d*=200. Dashed lines represent the mean values.

However, the distribution of the 20 NDNs (Fig. 3) shows that this is hardly possible because, on the one hand, the amounts of true and false DDAs overlap strongly and, on the other hand, there is a high variance in the similarity values. In addition, the values change depending on the size of the corpus [17]. To tackle the problems mentioned above, we present a new approach that allows us to estimate thresholds per drug entity independent of similarity values. Here we use information that we generate from the corpora and the results of our last experiment by showing that with a larger corpus (for all window sizes) the AVG precision also increases and so does the number of useful DDAs. *Is this also true for single entities? Does the precision and thus the number of meaningful DDAs also increase with an increasing amount of data per entity?* To answer these questions we have decided to investigate this possible correlation between Precision and 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 contained in the abstracts. Here, we count only DDAs which are present in at least 3 abstracts.

Evaluation Implementation. For our next experiment we sorted the two data set lists per drug-entity in ascending order and divided them into 10 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 conduct the experiment on the four corpora already introduced: 1900-1988, 1900-1998, 1900-2008, 1900-2018.

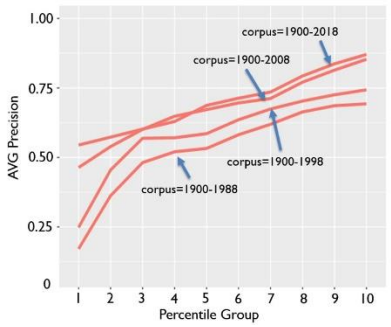

Figure 4. AVG precision values reached for different corpora and percentile groups with Document-Count.

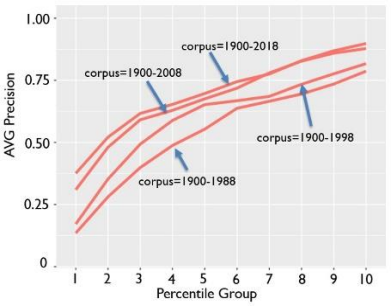

Figure 5. AVG precision values reached for different corpora and percentile groups with DDA-Count.

What we can see is that with increasing amounts of data (higher group) as well as in all corpora the precision increases (Figure 4, 5). Therefore, the more data available per drug-entity, the more

meaningful DDAs can probably be extracted. For the corpora 1900-2008 and 1900-2018 the values are almost the same. Thus, from a certain amount of data there seems to be no substantial improvement in precision. The curves for the DDA-Count (Figure 5) seem to be more similar in shape and course than the Document-Count curves (Figure 4). This could indicate that DDA-Count data correlation is more stable over time 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.

3.8 Correlation and Correlation-Stability

How can we calculate possible thresholds based on the previous results? 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. Therefore, we will investigate the following: At a fixed, "*meaningful*" precision value, how much more recall and predictions can we achieve compared to conventional methods such as k-NDNs or similarity thresholds?

How do we choose a *meaningful* precision value for our further investigations? If we set the precision p too high ($p > 95\%$) we have less 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 ~67%) of the DDAs is true - each DDA occurs in at least 3 abstracts - then there is a probability that the remaining minority is also meaningful and represents DDA predictions.

This remaining minority represents a kind of a *prediction buffer*. The importance of such a buffer can be demonstrated by the active substance *etoposide*, which is used in antitumor therapy. If we select the 20-NDNs for *etoposide*, then initially approx. 60% of the DDAs in the period 1900-1988 are true DDAs, whereby this value rises to 92% in the future (1900-2018). In our collection such a change applies to many drugs.

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 *true*-DDA precision of ~67%? And 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 classifier and for a given amount of data d of a drug x, to determine the number k of NDNs leading probably to the desired *meaningful* 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. In addition, we train the models with 200 dimensions, with the best window size (w=50) and the worst choice (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). As shown in Table 4, there is always a strong correlation exclusively with a window size of 50, together with a DDA-Count. In general, the correlation with a DDA-Count is always stronger than with a Document-Count. This indicates that the DDAs contained in texts have a stronger influence on contextualization. Therefore, it would be probably advantageous to use only documents with contained DDAs for training. Surprisingly, the larger windows size also seems to lead to generally more stable correlation values. We have the smallest standard deviation for w=50 (SD=0.018). Therefore, there is a chance that we can train a linear classifier based on the DDA-Count and at a window size of 50 and this independent of corpus size.

Table 4. PCCs for different window sizes and data counts. Best values in bold (SD=Standard Deviation).

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

3.9 Threshold Accuracy

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 certain similarity value. 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 time 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 linear classifier 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 a 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 5).

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 a 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%.

Our results are presented in Table 5. As it can be seen our approach leads to improved recall values. On average, the recall increases about 61% compared to Similarity-Threshold and 142% compared to a fixed-k-NDNs approach. Thus, we can extract more meaningful DDAs from the embedded space. The recall is *relatively* stable for the different time-periods and corpora. Thus, we can rely to a certain value-stability.

Table 5. Recall values for the different approaches and corpora. 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

4 TOWARDS CONTEXTUALISED NETWORKS

In the previous sections we addressed the question of *when* an embedded DDA is likely to be meaningful. The insights gained enable us to address the question of *why* two embedded entities are likely to be positioned close to each other in embedded space. If we can answer this question, we will be able to interpret and explore learned interrelationships between pharmaceutical entities.

We have described these (*difficult*) to interpret relationships in [9]. Here, we have grouped active substances by means of Neuronal Language Models, and we were able to show that these groupings are for the most part meaningful. However, in some cases the relationship seemed to be completely unclear. Since drugs have a strong link to diseases, embedded diseases may help to reveal latent links between embedded drugs. Thus, with our presented method we have created a possibility to extract variable sets of meaningful *k*-Nearest-Disease-Neighbors. An intersection of these sets probably represents a meaningful disease context between drug entities (Figure 6). This disease context could help us to reveal the hidden intrinsic relationship between different drugs.

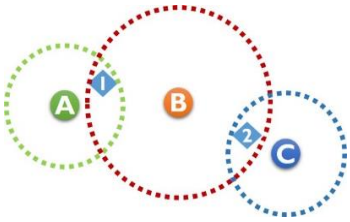

Figure 6. Simplified embedded space for drugs *A, B, C* and diseases *1, 2*. Dashed lines represent the area of meaningful k-NDNs.

In addition, we have an opportunity to reveal a latent contextualized network contained in literature with this linking-approach. For example, in our simplified model (Figure 6) we can link the drugs *A* and *B* using the disease *1* and we can link drugs *B* and C using disease *2*.

5 CONCLUSIONS AND FUTURE WORK

State-of-the-art approaches in the field of Neural Language Models (NLMs) enable progress in learning and predicting entity relations in terms of semantic quality, scalability, and performance. This is especially useful for automatically revealing new relationships between entities in literature. However, the neural component often leads to the question *why* a neural network has learned something. Answering this question becomes important if we want to use such models in scientific digital libraries to develop innovative services beyond term based searches. Because, if it is not clear *when* and *why* context led to a certain result, such models as Word2Vec may be rather unreliable in scientific literature-based exploration and discovery. Thus, extensive *whitebox-testing*, oriented towards specific semantic tasks, is a price we have to pay for possible disruptive approaches in the academic field and for academic searches. In this paper we have come closer to an answer to this question for important pharmaceutical entity relations.

Important hyper-parameters, such as window size and number of dimensions, as well as corpus properties, such as corpus size, are usually the first objects to be investigated in relation to the (semantic) analysis of NLMs. However, these are mostly specific to a given semantic task, as in our case in learning Drug-Disease Associations (DDAs). In this context we first examined the influence of window size on the learning of DDAs. Compared to the standard window size of 5, with a window size of 50 we can achieve a precision gain of 51% (for CDT) and 28% (for a co-occurrence approach). With increasing window size we can extract more meaningful DDAs. We could not achieve a comparable result with a different number of dimensions. Since the number of dimensions had hardly any effect on the learning of DDAs, we concentrated our investigation on the influence of window size in connection with corpus size. We divided our corpus into four corpora on different time periods. We were able to show that in every time period and corpus the following applies: With increasing window size the number of DDA predictions increases as well as the general number of meaningful DDAs. In addition, we have shown that the results get better with increasing corpus size. This led us to the hypothesis that the number of meaningful DDAs should also increase with increasing data volume per drug entity, such as Document Count and DDA-Count. In our experiments we have shown that as the amount of data per drug entity increases, the amount of meaningful embedded DDAs also increases. In short, for some drug-entities with larger data amounts, we should select more DDAs from the embedding space, and for drug-entities with smaller data amounts, fewer DDAs should be selected.

This led us to the question whether we could determine a *k*-Next-Disease-Neighbors (*k*-NDNs) threshold per drug entity on the basis of the data volume. If so, then there must be a (strong) correlation between the data sets and the DDA result. We have demonstrated this strong correlation (AVG Pearson Correlation Coefficient of 0.74) for the DDA-Count. This correlation is stable for different corpora and time periods (standard deviation 0.018). Therefore, by counting the DDAs occurring in the texts per drug-

entity, we can moderately predict how many k-NDNs we should extract per drug-entity from the embedding space, and this stays relatively stable for different corpora sizes and time-periods. This correlation also suggests that the documents containing a DDA are more important for relation-learning than those containing the drug alone. Based on our findings, we trained a linear *k*-NDN-threshold classifier and compared it with a Similarity-Threshold and a fixed-*k*-NDN-threshold approach. In all cases, our approach delivers improved results over all time periods, e.g. an improved AVG recall of min. 61% (compared to Similarity-Threshold) and max. 142% (compared to a fixed-*k*-NDNs-Threshold).

In our future work we want to investigate the properties of the proposed contextualized network and extend our presented methods to drug targets. Drug targets play a crucial role in important tasks like drug repurposing and the hope is that our literature based network will reveal new and previously unknown links.

ACKNOWLEDGMENTS

We thank the German Research Foundation (DFG) for the support of our project: *PubPharm - Specialized Information System Pharmacy* (Gepris 267140244). We thank Knut Baumann, Stephan Scherneck, Stefan Wulle, and Christina Draheim for their excellent feedback. Drug-centric web-services based on our research [9, 17] can be tested in the pharmaceutical digital library PubPharm (https://test.pubpharm.de/vufind/).

REFERENCES

[1] Gefen, D., Miller, J., Armstrong, J. K., Cornelius, F. H., Robertson, N., Smith-McLallen, A., & Taylor, J. A. (2018). Identifying patterns in medical records through latent semantic analysis. *Communications of the ACM*, *61*(6), 72-77.

[2] Owens, T. (2018, May). We Have Interesting Problems: Some Applied Grand Challenges from Digital Libraries, Archives and Museums. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (pp. 1-1). ACM.

[3] 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.

[4] Dumais, S.T. (2004). Latent Semantic Analysis. In *Annual review of information science and technology (ARIST)*, Vol. 38(1), Association for Information Science & Technology.

[5] 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.

[6] 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).

[7] Gottlieb, A., Stein, G. Y., Ruppin, E., & Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, *7*(1), 496.

[8] 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.

[9] Wawrzinek, J., & Balke, W. T. (2017, November). Semantic Facettation in Pharmaceutical Collections Using Deep Learning for Active Substance Contextualization. In *International Conference on Asian Digital Libraries* (pp. 41-53). Springer, Cham.

[10] Agarwal, P., & Searls, D. B. (2009). Can literature analysis identify innovation drivers in drug discovery? *Nature Reviews Drug Discovery*, *8*(11), 865.

[11] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[12] Dudley, J. T., Deshpande, T., & Butte, A. J. (2011). Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics*, *12*(4), 303-311.

[13] Weng, L., Zhang, L., Peng, Y., & Huang, R. S. (2013). Pharmacogenetics and pharmacogenomics: a bridge to individualized cancer therapy. *Pharmacogenomics*, *14*(3), 315-324.

[14] Agarwal, P., & Searls, D. B. (2009). Can literature analysis identify innovation drivers in drug discovery? *Nature Reviews Drug Discovery*, *8*(11), 865.

[15] 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.

[16] 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.

[17] 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.

[18] Mathur, S., & Dinakarpandian, D. (2012). Finding disease similarity based on implicit semantic similarity. *Journal of biomedical informatics*, *45*(2), 363-371.

[19] Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665-695.

[20] Hill, F., Kiela, D., & Korhonen, A. (2013). Concreteness and Corpora: A Theoretical and Practical Study. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)* (pp. 75-83).

[21] Chiu, B., Korhonen, A., & Pyysalo, S. (2016). Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 1-6).

[22] Rekabsaz, N., Lupu, M., & Hanbury, A. (2017, April). Exploration of a threshold for similarity based on uncertainty in word embedding. In *European Conference on Information Retrieval* (pp. 396-409). Springer, Cham.

[23] Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

[24] 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.

[25] 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*.

[26] 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.

[27] Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009, May). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19-27). Association for Computational Linguistics.

[28] Kiela, D., & Clark, S. (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)* (pp. 21-30).

[29] Chiu, B., Crichton, G., Korhonen, A., & Pyysalo, S. (2016). How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing* (pp. 166-174).

[30] Yin, Z., & Shen, Y. (2018). On the dimensionality of word embedding. In *Advances in Neural Information Processing Systems* (pp. 894-905).

[31] Grbovic, M., Djuric, N., Radosavljevic, V., Silvestri, F., & Bhamidipati, N. (2015, August). Context-and content-aware embeddings for query rewriting in sponsored search. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 383-392). ACM.

[32] Vulić, I., & Moens, M. F. (2015, August). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 363-372). ACM.

[33] Chiu, J. P., & Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, *4*, 357-370.

[34] 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, Cham.

[35] 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).