Semantic Facettation in Pharmaceutical Collections using Deep Learning for Active Substance Contextualization

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. Alternative access paths to literature beyond mere keyword or bibliographic search are a major success factor in today's digital libraries. Especially in the sciences, users are in dire need of complex knowledge spaces and facettations where entities like e.g., chemical substances, genes, or mathematical formulae may play a central role. However, even for clear-cut entities the requirements in terms of contextualized similarities or rankings may strongly differ. In this paper, we show how deep learning techniques used on scientific corpora lead to a strongly contextualized description of entities. As application case we take pharmaceutical entities in the form of small molecules and demonstrate how their learned contexts and profiles reflect their actual use as well as possible new uses, e.g., for drug design or repurposing. As our evaluation shows, the results gained are quite comparable to expensive manually maintained classifications in the field. Since our techniques only rely on deep embeddings of textual documents, our methodology promises to be generalizable to other use cases, too.

Keywords: Digital libraries, information extraction, facettation, deep learning.

1 Introduction

In pharmaceutical digital libraries, (active) substance similarity forms the basis of various innovative services for information access such as structure search, grouping and facettation of drugs, suggestion lists and many others. However, what makes a similarity measure between entities semantically *meaningful* in a domain? While there usually is no single universally true answer, there are generally several accepted methods of determining similarity differing in their complexity, accuracy, and applicability given a task. Yet, from a digital library provider's perspective, there is another important distinction between these similarity measures: can the necessary features for their computation be extracted automatically in a scalable way or are they based on semantic features that still need expensive manual curation? Given the current promising developments in automatic information extraction and the indexing challenges posed by rapidly increasing publication numbers, this is indeed a central question.

Consider the example domain of pharmaceutical collections: Here, to compute active substance similarity two approaches are widely used: on one hand a (sub-)structural

similarity (chemical or molecular similarity), and on the other hand a taxonomical similarity regarding therapeutic uses, etc. (usually curated manually by domain experts).

For efficiently deriving *(sub-)structural similarity* between substances, all molecular structures are usually encoded in bit-string fingerprints. To reduce dimensionality and ease comparison the bits are set with respect to molecular features such as atom sequences, ring compositions or atom pairs of each molecule. The exact composition of fingerprints may thus vary depending on the specific use case and research field [1]. However, this does not only result in numerous and different fingerprint types (e.g. Extended Fingerprint, MACCS, Estate, etc.), but also in different similarity measures between substances, such as Tanimoto, cosine, dice, etc. In brief, the combination of fingerprints and similarity measures leads to a wide variety of possible results, and it is interesting to note that their respectively induced rankings of most similar substances are usually only weakly, if at all correlated [2]. Moreover, while structural similarity is extremely useful for screening, it does not capture other important semantic features.

The *taxonomical similarity* approach to compute active ingredient similarity is based on mostly manually curated semantic classification systems. Drugs, chemicals, or in general active ingredients are grouped according to their chemical, therapeutic or anatomical features. Considering pharmacy, there are a couple of popular classification systems such as the Medical Subject Headings (MeSH) Trees[1], the Anatomical Therapeutic Chemical (ATC) Classification System[2] or the American Hospital Formulary Service (AHFS) Pharmacologic-Therapeutic classification[3]. Of course, their applicability is limited by the actual number of substances indexed: querying DrugBank[4] as a relatively complete resource [3], most active ingredients are not classified by any of the above-mentioned classification systems.

Recently, many research efforts have considered a new way of generating semantically meaningful similarities for scientific entities: *facettation with categories dynamically created from large document corpora* (for a good overview see [4]). Indeed, the enrichment of entity metadata with information from different sources like external knowledge bases or focused document collections has been proven extremely successful in scientific search scenarios, see e.g., [5] and [6]. The key to success can be seen in a *contextualization of entities* as expressed by their actual use in research, which is in turn reflected in respective publications. In this paper, we present a novel deep learning-based technique to contextualize entities. Following our pharmaceutical use case, we evaluate our method over the PubMed collection and show that the facets gained from embeddings in high-dimensional document spaces are semantically meaningful, while measuring similarity regarding different entity aspects. Thus, our method adds alternative facets statistically justified by a large body of existing research publications, giving users easy access to hidden entity semantics for digital library searches. Moreover, these facets can be automatically derived without expensive manual curation.

---

1 https://www.nlm.nih.gov/mesh/intro_trees.html
2 https://www.whocc.no/atc_ddd_index/
3 http://www.ahfsdruginformation.com/ahfs-pharmacologic-therapeutic-classification/
4 https://www.drugbank.ca/

The paper is organized as follows: section 2 revisits related work. Section 3 details our method for facettation of drugs, accompanied by an extensive evaluation against curated classification systems in section 4. We close with conclusions in section 5.

2 Related Work

Capturing semantically meaningful similarities for scientific entities has since long been an active field of research. Today, most recognized systems are to a large degree still *manually maintained* to guarantee usage experience and to provide a reliable foundation for value adding services and research planning. While the current explosion of scientific results clearly calls for automation, the quality of resources cannot be compromised, i.e. a high degree of precision has to be maintained. The most prominent classification systems (later used as ground truth) for pharmaceutical uses are:

- The *Anatomical Therapeutic Chemical (ATC) Classification System*. ATC subdivides drugs according to their therapeutic uses and chemical features. Maintained by the World Health Organization (WHO), it is currently the most used drug classification system and serves as an important source for tasks like e.g., drug repurposing and drug therapy composition [7].
- The *Medical Subject Headings (MeSH)*. MeSH is a controlled vocabulary and serves as general classification system for biomedical documents in MEDLINE maintained by the National Library of Medicine (NLM). MeSH descriptors are organized in 16 main categories, e.g. category C for diseases and D for drugs, further divided in finer levels (subgroups) leading to a hierarchical structure.
- The *American Hospital Formulary Service (AHFS)*. AHFS distinguishes drugs according to their pharmacologic and therapeutic effect with a focus on drug therapies. Like ATC and MeSH, AHFS shows a hierarchical structure.

Manual drug annotation may yield superior quality, but it is also related with high costs. Therefore, in recent years many approaches to *annotate drugs automatically* have been designed. In general, these approaches rely on a blend of machine learning, information retrieval, and information extraction techniques. To annotate properties in pharmaceutical texts reliably, a wide variety of methods has been devised. For instance, [7] employs support vector machines to predict ATC class labels for yet unclassified drugs and shows that given rich training sets, document-based classification can actually outperform classifications performed on chemical structures only. For the same task, [8] shows the power of text mining to create enriched drug fingerprints and after some manual curation their subsequent benefit for retrieval. In [9] an approach for the automatic annotation of biomedical documents with MeSH terms is presented. Different classification systems are compared to reproduce manual MeSH annotations.

With classification accuracies of already around 80%, all of the above document-based approaches show the benefits and general applicability of text mining for entity metadata enrichment. Thus, a domain-specific contextualization of entities in scientific digital libraries seems appealing. To find central topics in documents two major approaches have been used: latent semantic analysis (LSA [10]) performs singular value

decompositions over term-document matrices to get topics as linear combinations of vocabulary terms. Latent Dirichlet Allocation (LDA [11]) sees documents as mixtures of different topics, where each term's generation is attributable to one of the document's topics. Since both models show problems in NLP tasks like polysemy detection or syntactic parsing, recently Word Embeddings [12] quantifying and categorizing semantic similarities between linguistic items based on their distributional properties in large samples of language data have been proposed as a powerful deep learning alternative. Therefore, in the following we will rely on word embeddings as the state of the art method for entity contextualization and in particular, will use the Word2vec Skip-Gram model implementation from the open source Deep-Learning-for-Java[5] library.

3 Building New Facets based on Word Embeddings

The basic idea of our approach is to create a new contextualized facet for entity-based search in scientific digital libraries: in particular, a selection of closely related entities with respect to the search entity. For actually building contextualized facets every corpus of scientific documents can be used, but normally the selection of the document base for subsequent embedding strictly reflects the type of entities under scrutiny. For example in the case of pharmaceutical entities such as active ingredients, the National Library of Medicine's PubMed collection would be a good candidate.

After the initial crawling step the following process can be roughly divided into four sub-steps:

1. *Preprocessing of crawled documents.* After the relevant documents were crawled, classical IR-style text preprocessing is needed, i.e. stop-word removal and stemming. The preprocessing helps mainly to reduce vocabulary size, which leads to an improved performance, as well as improved accuracy. Due to their low discriminating power, all words occurring in more than 50% of the documents are removed. These are primarily often used words in general texts such as '*the*' or '*and*', as well as terms used frequently within a domain (as expressed by the document base), e.g., '*experiment*', '*molecule'*, or '*cell*' in biology. Stemming further reduces the vocabulary size by unifying all flections of terms. A variety of stemmers for different applications is readily available.

2. Creating word embeddings for entity contextualization. Currently, word embeddings [12] are the state-of-the-art deep learning technique to map terms into a multi-dimensional space (usually about 200-400 dimensions are created), such that terms sharing the same context are grouped more closely. According to the distributional hypothesis, terms sharing the same context in larger samples of language data quite often, in general also share similar semantics (i.e. have similar meaning). In this sense, word embeddings group entities sharing the same context and thus collecting the nearest embeddings of some search entity leads to a group of entities sharing similar semantics.

3. Filtering according to entity types. The computed word embeddings comprise at this point a large portion of the corpus vocabulary. This means, for each vocabulary term there is exactly one word vector representation as output of the previous step. Each

[5] https://deeplearning4j.org/

vector representation starts with the term followed by individual values for each dimension. In contrast, classical facets only display information of the same type, such as publication venues, (co-)authors, or related entities like genes or enzymes. Thus, for the actual building of facets, we only vector representations of the same entity type are needed. Here, dictionaries are needed to sort through the vocabulary for each type of entity separately. The dictionaries either can be directly gained from domain ontologies, like e.g. MeSH for illnesses, can be identified by named entity recognizers like e.g., the Open Source Chemistry Analysis Routines (OSCAR, see [13]) for chemical entities, or can be extracted from open collections in the domain, like the DrugBank for drugs.

4. Clustering entity vector representations. The last step is preparing the actual facettation of entities closely related to some search entity. To do this, we first consolidate the individual document spaces of the filtered entities by multidimensional scaling (reducing its dimensionality to about 100-150). This steep dimensionality reduction removes noise and enables a meaningful subsequent clustering. We then apply a k-means clustering technique on all representations and decide for good cluster sizes: in our approach optimal cluster sizes are not decided by a fixed threshold, but by an analysis of intra-cluster vs. inter-cluster similarity.

While the basic algorithm promises to be applicable for a wide variety of domains, testing its effectiveness in creating high quality entity facets needs a domain specific focus. The following section evaluates our approach in a pharmaceutical use case.

4 Evaluation of Entity Contextualization

For the evaluation, we will first describe our pharmaceutical text corpus and basic experimental set-up decisions. Moreover, we perform a ground truth comparison and show the meaningfulness of the facets automatically derived by our facettation method: we compare results with the three established classification systems from section 2.

4.1 Experimental Setup and Algorithm Implementation

Experimental Setup.

Evaluation corpus. With more than 27 million document citations, *PubMed*[6] is the largest and most comprehensive digital library in the biomedical field. However, since many documents citations do not feature full texts, we relied solely on abstracts for learning purposes. As an intuition, the number of abstracts matching each pharmaceutical entity under consideration should be 'high enough' because with more training data, contexts that are more accurate can be learned, yet the computational complexity grows. Thus, we decided to use the 1000 most relevant abstracts for each entity according to the relevance weighting of PubMed's search engine.

Query Entities. As query entities for the evaluation, we randomly selected 275 drugs from the *DrugBank*[7] collection. We ensured that each selected drug featured at least

[6] https://www.ncbi.nlm.nih.gov/pubmed/
[7] https://www.drugbank.ca/

one class label in ATC, MeSH, or AHFS, and occurred in at least 1000 abstracts on PubMed. Thus, our final document set for evaluation contained 275.000 abstracts. As ground truth, all class labels were crawled from both, *DrugBank* and the *MeSH thesaurus*[8]. For example, all retrieved classes for the drug '*Acyclovir*' are shown in Table 1. Since all classification systems show a too fine-grained hierarchical structure, we remove all finer levels before assigning the respective class label to each drug. For example, one of the ATC classes for the drug 'Acyclovir' is '*D06BB53*'. The first letter indicates the anatomical main group, where '*D*' stands for 'dermatologicals'. The next level consists of two digits '*06*' expressing the therapeutic subgroup 'antibiotics and chemotherapeutics for dermatological use'. Each further level classifies the object even more precisely, until the finest level usually uniquely identifies a drug.

Table 1. Classes in different classification systems for the drug 'Acyclovir'.

Classification System	Assigned Classes
ATC	J05AB01, D06BB53, D06BB03, S01AD03
AHFS	08:18.32, 84:04.06
MeSH Trees	D03.633.100.759.758.399.454.250

Algorithm implementation and parameter settings.

1. Text Preprocessing: Stemming and stop-word removal was performed using a *Lucene*[9] index. For stemming we used Lucene's *Porter Stemmer* implementation.

2. Word Embeddings: After preprocessing, word embeddings were created with DeepLearning4J's *Word2Vec*[10] implementation. To train the neural network, we used a minimum word frequency of 5 occurrences. We set the word window size to 20 and the layer size to 200 features per word. Training iterations were set to 4. We tested several settings, but the above-mentioned turned out best for subsequent clustering.

3. Entity filtering. While Word2Vec generated a comprehensive list of word vector representations, we subsequently filtered out all vectors not related to any DrugBank entity (resulting in 275 entity-vectors). For corpus consolidation (dimensionality reduction) after the filtering step, we used a *Multidimensional Scaling* (MDS[11]) technique: we scaled word vector representations from 200 down to 120 dimensions. The intention of the MDS step was to smooth out possible noise. Smoothing out noise in high-dimensional representations can have a positive impact on overall performance [15]. Whereby overall performance means in our case an improvement in F-score. Compared to unscaled entity-vectors, the MDS step resulted in an improvement of ~10% in F1-score. In addition, we tested the MDS with different parameters, with respect to F1-score best results were achieved with a scaling to 120 dimensions. Surprisingly, an initial layer size setting of 120 features (for Word2Vec training) did not lead to a similar improvement. Instead the result was comparable to results achieved with a layer size setting of

[8] https://meshb.nlm.nih.gov/search
[9] https://lucene.apache.org/
[10] https://deeplearning4j.org/word2vec
[11] http://algo.uni-konstanz.de/software/mdsj/

200 features but without an additional MDS step. We conclude that the improvement in F1-score is the consequence of the MDS step.

For the MDS step, we also experimented with different similarity measures to calculate the dissimilarity matrix: best results were achieved using cosine similarity to calculate the matrix.

4. Clustering vector representations. In this step, we clustered the 275 entity vector representations obtained in the previous filtering step. For the clustering step we used Apache Commons' Multi-KMeans[12]++ clusterer. For a fair comparison to our ground truth, our goal is to choose the class most suitable for a drug as well as for the entire cluster. Thus, for comparing class labels of entities within a cluster, we assign the majority class label to each cluster and regard all entities in that cluster sharing the majority label as true positives. To avoid double counting these true positives as false positives for additional labels they carry, we strip all remaining class labels. Entities in a class not sharing the majority class label are false positives and will be labeled with their respective label that is most frequent in that class. Again, to avoid double counting all other labels are removed.

4.2 Experimental Evaluation

For the experimental evaluation, we first have to determine what quality criteria a document-centric contextualization approach should meet to be useful for dynamically creating entity facets. Since the subsequent facettation will be based on the clusters generated by our approach (i.e. for each query entity all other entities sharing its cluster will be presented in the facet), each cluster has to exhibit certain criteria:

- *Semantic accuracy*: A facet should group entities under some *common theme* that seems most suitable with respect to the query entity. This is influenced by the semantic purity of clusters as well as a good trade-off between precision and recall. Since higher recall values might produce overly large facets, the emphasis should rather be on reaching higher precision values.
- *Semantic coverage*: For a good handling of the subsequent facets, the distribution of entities over the clusters should be *well balanced*. Clusterings exhibiting many large and/or many small clusters will result in unsatisfactory usage experience in the respective faceted interface.
- *Semantic suitability*: The selected entities per facet should be *clearly justified* by the underlying document collection. Since there are different document-centered approaches, a quantitative comparison regarding a ground truth is needed.

Semantic Accuracy of the Facettation: In our first experiment, we test the semantic accuracy of our facettation, i.e. how well do entities in each cluster reflect a common topic. Since this is obviously dependent on cluster sizes (smaller clusters inherently show higher purity) and the respective granularity of the topic (in the sense of semantic distances), we will vary both, the *number of clusters* in the clustering procedure and the *granularity* of the topics (first level vs. second level accuracy). As ground truth, we use

[12] http://commons.apache.org/proper/commons-math/

only the categories given by the largest three pharmaceutical classification systems ATC, MeSH, and AHFS (see section 2). Please note that this ground truth restriction is overly strict on document-centered contextualization, since commonly understood contexts reflected in literature might not be reflected by any of the three systems. Thus, our experiments can be seen as a worst-case boundary for our approach.

First, we quantify the accuracy in terms of precision/recall and F-measures on the top categorization level only. We use the standard method for clustering accuracy described in [14]. Because facets should tend towards higher precision for improved user experience, we report both, F1- and F0.5-scores. We vary the number of clusters (k) in our k-means clustering between 10 and 80. Since the randomly chosen query entities might not be evenly distributed over the respective categories chosen as majority labels, we compare our approach against a base line of clusters, where items have been randomly exchanged between clusters. If there would be clearly dominant categories, such a random baseline would show high accuracies.

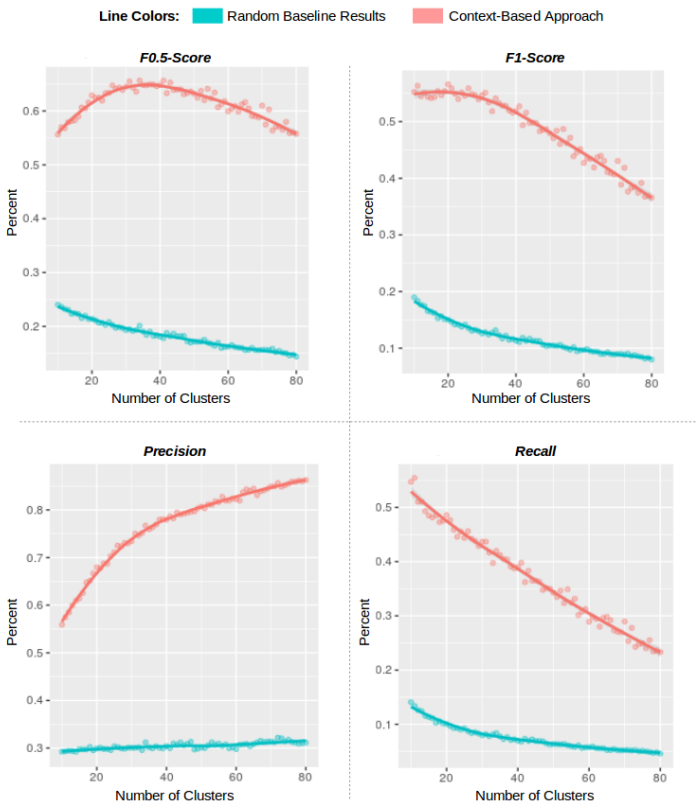

Fig. 1. Comparison of contextualized facettation (red) and random clustering (blue).

Figure 1 shows averaged results of 30 independent runs for each number of clusters. As could be expected, precision steeply increases for higher numbers of clusters (i.e.

small cluster sizes), whereas recall decreases the more clusters are built. However, the F-scores show a clear optimum at 25 clusters (F1-score) and 35 clusters (F0.5-score). Hence, preferring smaller cluster sizes (on average of 8-10 entities per facet) in stark contrast to the random baseline that always prefers the smallest number of clusters possible. Moreover, our approach's F-scores constantly outperform the baselines with 0.55 (F1-score) and 0.65 (F0.5-score) reaching precisions beyond 80%. Thus, surprisingly our generalist approach is even comparable in overall accuracy to approaches specifically designed to predict ATC or MeSH classifications, as reported in section 2.

We repeated the above experiments for the second layer of granularity in the classification systems and achieved quite similar results (graphs have been omitted for space reasons), again clearly outperforming the baseline. Of course, with finer granularity the relative size of clusters has to be expected to be much lower. However, again measuring the F0.5-score, we achieved best results with a moderate 97 clusters at an accuracy level of still 0.61. This is only 4% less, compared to the first level of granularity. For the F1-score, best results were achieved with 69 clusters at an accuracy level of 0.55.

Semantic coverage of the Facettation: To investigate how well the individual semantics of the different categorization systems are reflected by our contextualized facets, we show that our facettation is indeed balanced, i.e. it does not generate extreme distributions in either cluster sizes or majority label provenance. For instance, it would not be desirable, if our facettation created one single big facet, while the remaining facets only contain a single entity each. Moreover, the distribution of majority class label regarding their respective source classification system should be balanced.

Again, we performed experiments on two levels of granularity: top-level and second level. For the top-level granularity we calculated average cluster sizes for the sweet spot (i.e. at $k = 35$ clusters) of our last experiment and show the respective results as box plots in Figure 2. As we can clearly see, there are only few larger clusters, while the majority of clusters features between 3 and 8 entities, with a median of 4.8. Clusters with sizes smaller than 3 are quite rare. Moreover, it is encouraging to note that the overall distribution of entities in clusters strongly resembles the distribution exhibited by the respective classification systems. That means, the cluster sizes decided by our deep learning-based contextualization are on the correct resolution level, which together with the high accuracy speaks for a good semantic coverage.

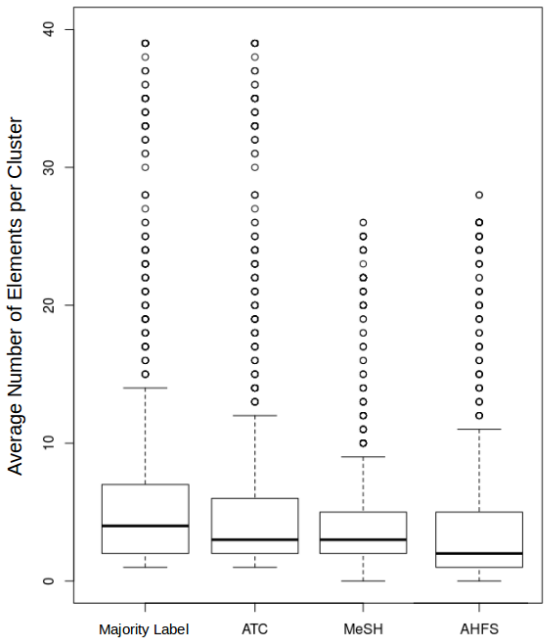

Fig. 2. Average cluster sizes on first level granularity for the majority label compared to ATC, MeSH, and AHFS.

On the second level of granularity (see Figure 3) the medians of the distributions are noticeably lower, as was to be expected for higher number of clusters (k = 97). Still, our approach's distribution again closely resembles the distributions of the respective classification system. Moreover, in contrast to MeSh and AHFS our approach avoids empty clusters and shows fewer outliers with large cluster sizes, quite similar to the ATC classification system.

Looking at the provenance of majority cluster labels we find that on top-level granularity the majority labels chosen for each cluster on average reflect 60.3 % from ATC classes, 34.3% from MeSh tree classes, and 5.4% from AHFS classes. For second level granularity, we get 51.8 % from ATC, 36.8% from MeSh, and 11.4% from AHFS. Thus, our contextualization approach does indeed reflect different semantics as given by the individual, manually created classification systems.

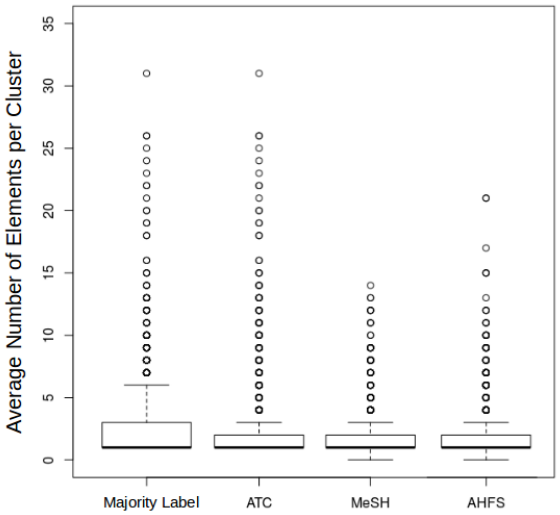

Fig. 3. Average cluster sizes on second level granularity for the majority label compared to ATC, MeSH, and AHFS.

Semantic suitability of the Facettation: In our last experiment, we compare the clustering accuracy of our approach with the accuracy achieved by classical IR techniques based on term frequencies. Hence, we computed a TF-IDF-weighted vector space model on all pharmaceutical texts in our selected document corpus for the 275 query entities, again followed by a *k*-means clustering step. We then compared the respective accuracies of the two methods with respect to the three manual classification systems as ground truth.

In the clustering step for the top-level granularity, also TF-IDF shows highest accuracy values for a number of 35 clusters and thus seems quite suitable for the task. However, in comparison with a TF-IDF-weighted vector space model, the contextualized facets achieved noticeable improvements with respect to accuracies: the F0.5-score was on average 30% higher, and the F1-score still 18% higher. In brief, our deep learning-based approach leads to a much higher precision as compared to classic IR-style frequency-based approaches.

5 Conclusions and Future Work

In this paper, we presented a novel deep learning-based technique to contextualize entities for building semantically meaningful facettations in pharmaceutical collections. In pharmaceutical digital libraries, substance similarity forms the basis for various innovative services for information access such as finding active ingredients or structure search. Today, substance similarity is based either on manually curated semantic clas-

sification systems, or on comparisons of the underlying chemical structures. Both methods are extremely useful, but on the one hand chemical structure approaches do not capture important semantic features, on the other hand most active ingredients are not classified by manually curated categorization systems.

We demonstrated in our experiments, that our proposed method for a new facettation of active ingredients, achieves a high semantic accuracy. Since, on both levels of granularity, our approach constantly outperforms the baselines as well as reaches high precisions (beyond 80%). Thus, our facettation method clusters active ingredients in a meaningful way and therefore elements, contained in the same facet, share with a high accuracy a similar semantic. Next, we proved the sematic coverage of the facettation by investigating how well the individual semantics of the different categorization systems are reflected by our contextualized facets. Here, on both levels of granularity the different majority labels are moderate distributed. Moderate means, none categorization type dominates the overall facettation. Thus, our contextualization approach does reflect different semantics as given by the individual, manually curated categorization systems. This in turn shows that a facet consist of a composition of different categorization systems, in which the facet elements (active ingredients) share a similar semantic. In our pharmaceutical case, the facettation can be a suitable alternative to expensive as well as in most cases incomplete manually curated categorization systems. Moreover, we also demonstrated that our facettation is balanced and does not generate extreme distributions cluster sizes. Since, small (cluster size < 3) as well as very large cluster are quite rare. Thus, it reflects a given distribution in respect to the different categorization systems and therefore facets have a similar size compared to manually curated categorization system categories. Finally, we tested the semantic suitability of the facettation by comparing it with classical IR techniques. Our approach outperformed (up to 30%) TF-IDF-weighted vector space model. Therefore, our deep learning-based approach is a suitable alternative for classic IR-style frequency-based approaches.

In addition to the statistical evaluation presented in this paper, we also questioned domain experts for a first interpretation of our facettation. Surprisingly, they found hidden semantics for some of the low-accuracy facets. This may indicate that our facettation technique is able to discover hidden active ingredient contexts. A better understanding of such hidden contexts would be interesting. Furthermore, labeling of facets was however not considered in this paper. Such a labeling would prove quite useful for an interpretation of the individual facets as well as it could lead to a better understanding with respect to our facettation.

References

1. Willett, P., Barnard, J. M., Downs, G. M. (1998). Chemical similarity searching. In *Journal of chemical information and computer sciences*, Vol. 38(6), 983-996.
2. Tönnies, S., Köhncke, B., Balke, W.T. (2011). Taking chemistry to the task: personalized queries for chemical digital libraries. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Ottawa, Canada.

3. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, J. (2006) *DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Research*, Vol. 1;34 (Database issue):D668-72.
4. Sacco, G.M., Tzitzikas, Y. (2009) *Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience*. Springer.
5. Köhncke, B., Balke, W.T. (2013). Context-Sensitive Ranking Using Cross-Domain Knowledge for Chemical Digital Libraries. In *International Conference on Theory and Practice of Digital Libraries (TPDL)*. Valletta, Malta.
6. Gonzalez Pinto, J. M., Balke, W.T. (2015). Demystifying the Semantics of Relevant Objects in Scholarly Collections: A Probabilistic Approach. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Knoxville, TN, USA.
7. Gurulingappa, H., Kolárik, C., Hofmann-Apitius, M., Fluck, J. (2009). Concept-based semi-automatic classification of drugs. *Journal of chemical information and modeling*, Vol. 49(8).
8. Dunkel, M., Günther, S., Ahmed, J., Wittig, B., Preissner, R. (2008). SuperPred: drug classification and target prediction. *Nucleic acids research*, *36*(suppl 2), W55-W59.
9. Trieschnigg, D., Pezik, P., Lee, V., De Jong, F., Kraaij, W., Rebholz-Schuhmann, D. (2009). MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*, Vol. 25(11), Oxford University Press.
10. Dumais, S.T. (2004). Latent Semantic Analysis. In *Annual review of information science and technology (ARIST)*, Vol. 38(1), Association for Information Science & Technology.
11. Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet Allocation. In *Journal of Machine Learning Research*, *3*(Jan), MIT Press.
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, USA.
13. Jessop, D.M., Adams, S.E., Willighagen, E.L., Hawizy, L., Murray-Rust, P. (2011) OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, Vol. 3(1), Springer.
14. Manning, C. D., Raghavan, P., Schütze, H. (2008). *Introduction to information retrieval.* Cambridge University Press.
15. Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.