Quality Assessment of Knowledge Graph Hierarchies using KG-BERT

Kinga Szarkowska[1], Véronique Moore[2], Pierre-Yves Vandenbussche[2] and Paul Groth[1]

[1]*University of Amsterdam, Science Park 904, 1098 XH Amsterdam, Netherlands*
[2]*Elsevier Amsterdam, Radarweg 29a, 1043 NX Amsterdam, Netherlands*

Abstract

Knowledge graphs in both public and corporate settings need to keep pace with the constantly growing amount of data being generated. It is, therefore, crucial to have automated solutions for assessing the quality of Knowledge Graphs, as manual curation quickly reaches its limits. This research proposes the use of KG-BERT for a triple (binary) classification task that assesses the quality of a Knowledge Graphs's hierarchical structure. The use of KG-BERT allows the textual as well structural aspects of a Knowledge Graph to be leverage for this quality assessment (QA) task. The performance of our proposed approach is measured using four different Knowledge Graphs: two branches (Physics and Mathematics) of a corporate Knowledge Graph - OmniScience, a WordNet subset, and the UMLS Semantic Network. Our method yields high-performance scores on all four KGs (88-92% accuracy) making it a relevant tool for quality assessment and knowledge graph maintenance.

Keywords

Ontology Maintenance, Knowledge Graphs, Hierarchical Knowledge Graphs, hierarchy evaluation, triple classification, contextual word embeddings, BERT, KG-BERT.

1. Introduction

Knowledge Graphs are at the heart of many data management systems nowadays, with applications ranging from knowledge-based information retrieval systems to topic recommendation [1]and have been adopted by many companies [1]. Our research originated with the need for the automatic quality assessment (QA) of OmniScience [2], Elsevier's cross-domain Knowledge Graph powering applications such as the Science Direct Topic Pages.[1]

A Knowledge Graph (KG) is a graph representation of knowledge, with entities, edges, and attributes [3]. Entities represent concepts, classes or things from the real world, edges represent the relationships between entities, and attributes define property-values for the entities. We refer to these sets as "triples".

A number of QA dimensions have been identified for KGs [4]. Here, we focus on the semantic accuracy dimension: the degree to which data values correctly represent the real-world facts (or concepts) [4]. More specifically, we focus on the hierarchy evaluation of KGs: whether

ISWC 2021: Deep Learning for Knowledge Graphs, October 24–28, 2021, Virtual Conference
kinga.szarkowska@gmail.com (K. Szarkowska); v.malaise@elsevier.com (V. Moore); p.vandenbussche@elsevier.com (P. Vandenbussche); p.t.groth@uva.nl (P. Groth)

[1]https://www.sciencedirect.com/topics/index

its hierarchical structure is correctly represented. To do this, we employ contextual word embeddings [5] and investigate the use KG-BERT method [6]'s binary classification task. It is a binary triple classification task, as for a given KG triple (entity, relation, entity/literal) the classifier will return whether a given triple is correct or not.

KG-BERT takes advantage of the textual representation of the data for assessing the veracity of KG triples and uses transfer learning to fine-tune embeddings pre-trained using the BERT model [5] for a triple classification task. *Our novel contribution is to use textual representations of the KG hierarchy in combination with KG-BERT to evaluate hierarchy quality.* We evaluate the performance of this method on four different KGs.

The paper is structured as follows: in the related work section, we present KGs evaluation frameworks and approaches for triple classification tasks. In the methodology section, we describe the datasets that were used for this research, along with our sampling strategy, a detailed description of the KG-BERT method and the evaluation metrics that we have selected. In the results section, we present results for the four selected KGs. Lastly, we summarize the outcomes of our research and propose future directions for exploration.

2. Related Work

In this section, we present KGs quality assurance frameworks and methods developed for KGs triple classification tasks.

2.1. Knowledge Graph Quality Assurance Frameworks

One of the first frameworks that has been widely used for the data quality assessment of a hierarchical triples structure was developed in 1996 by Wang and Strong [7]. They identified four main quality dimensions: intrinsic, contextual, representational, and accessibility [7]. A framework proposed in 2007 [8] and built on [7] introduces a taxonomy of Information Quality dimensions. Zaveri et al. [4] focused on a quality evaluation framework for linked data, gathering 18 quality dimensions (with 69 quality metrics), including the dimensions introduced by [7]. Chen et al. [9] adjusted the framework proposed by [4] for KGs. They created 18 requirements on knowledge graphs quality and mapped them to knowledge graph applications. Raad and Cruz [10] also gathered evaluation methodologies for ontologies. Raad and Cruz distinguished several validation criteria (such as accuracy, completeness, consistency) and presented four evaluation approaches.

The dimension that maps to the evaluation (assessment) of automatically expanding large scale KGs is the semantic accuracy dimension [4], or just accuracy [10]. Zaveri et al. [4] define it as the degree to which data values correctly represent real-world concepts. Our research develops a method to evaluate one component of this dimension: whether the hierarchy of a KG is accurate.

2.2. Knowledge Graph Triple Classification

We propose to use a (binary) triple classification task approach to address the problem of KG hierarchy evaluation. We describe how our method compares with the current state-of-the-art

in KG evaluation.

A triple classification task aims to predict whether an unseen Knowledge Graph triple is true or not. For that we can consider KGs entities and their relationships as real-valued vectors, and assess the triples plausibility. Real-valued vector representations are called Knowledge Graph embeddings [11]. There are two types of approaches for creating the KG embeddings: translational distance models and semantic matching models [11]. The main difference is the use of scoring function: translational models use distance-based functions and semantic matching models use similarity-based scoring functions. Examples of translational models are: TransE [12], TransH [13], TransR [14], or TransG [15]. RESCAL [16] and DistMult [17] are representatives of semantic matching models. All of these methods use only the data's structural information, and do not leverage external information such as entity type, or textual description, to improve model performance.

KG-BERT was introduced to address this limitation by taking advantage of textual information in addition to structural information [6]: instead of creating Knowledge Graph embeddings to represent the structure of the data (using the relationships presented in the data), it uses a textual (distributed) representation of the triples in a corpus. It first creates a representations of each of the entities in a triple using pre-trained BERT [5] embeddings. These embeddings are then updated to minimize the loss function for the triple classification task. Finally, a binary classifier predicts from its embedding, whether a triple is true or not with an associated confidence score.

One consequence of initialization through BERT is that the model takes advantage of knowledge encoded within BERT during its training. This feature was one of the reasons to explore the use of KG-BERT for the this evaluation task.

3. Methodology

We elaborate now on our technical and methodological decisions. First, we present the datasets that are used in this research. Each dataset is from a different domain and used in a different community, but they all hierarchically structured. We introduce our negative sampling approaches as we evaluate our results against a gold set of triples that should only be true. We then describe KG-BERT in more detail, along with the evaluation approach.

3.1. Datasets

We selected four Knowledge Graphs from different domains (described in more details in Table 1), which are all hierarchical knowledge graphs (structured with the classification inclusion [18] relation). We discarded here the benchmark datasets of WN11 and FB13 typically used for triple classification tasks, as they contain a wide variety of relationships and were not representative of our target KGs.

WordNet [19] is a large lexical database representing relationships between words in the English language. It is widely used for tasks such as natural language processing or image classification [20]. We extracted a subset of WordNet focused on the classification inclusion

Table 1
Overview of the experimental datasets

Dataset	Example of the Triple	No. of Triples	No. of Entities
WordNet	*Refried Beans* is a *Dish*	62,323	48,048
UMLS	*Enzyme* is a *Chemical*	500	135
Physics	*Seismology* is a concept of *Geophysics*	5,404	3,697
Mathematics	*Outlier* is a concept of *Summary Statistics*	2,770	2,126

relation: we extracted only nouns - excluding proper names - in a hyponymy relation. We performed this filtering on the hyponymy Wordnet subset.[2]

The Unified Medical Language System (UMLS) [21] is a comprehensive ontology of biomedical concepts. UMLS is made of two distinct KGs: the Semantic Network and a Metathesaurus of biomedical vocabularies. The Semantic Network represents broad subject categories together with their semantic relations. It groups and organizes the concepts from the Metathesaurus in one of the four relationships: "consist of", "is a", "method of", "part of".

We selected the Semantic Network for our research because it has more generic concepts than the Metathesaurus and is better structured. We decided to investigate the performance of our model on the subset of that network consisting of the triples in a hierarchical ("is a") relationship (the classification inclusion relation).

OmniScience [2] is an all-science domains knowledge graph that is used, amongst others, for the annotations of research articles. It connects external publicly available vocabularies with the entities required in Elsevier platforms and tools. It is maintained by scientific experts from Elsevier. OmniScience is a poly-hierarchy, in which scientific concepts can belong to multiple domains. The relationship between OmniScience concepts can be described as a hyponymy relation, or "is a" relation. OmniScience has several domain branches, such as Physics, Mathematics, or Medicine and Dentistry. We used two branches as test cases, namely Physics and Mathematics.

3.2. Negative Sampling

We considered all of the KGs above to consist of correct triples. Therefore, negative sampling is necessary to prepare a training set for a classification problem. We followed the approach proposed in [22] and use the 1:3 ratio for negative samples. We followed three strategies to generate these negative samples:

1. per each head entity we randomly sample a tail entity;
2. per each tail entity we randomly sample a head entity;
3. per each pair we exchange the head entity with the tail entity, which gives us “reversed” samples, that should help train the model with respect to the direction of the relation.

After sampling 3 negative examples per each proper pair, we filtered out all of the generated samples did that occur in the original set of triples from KG to ensure that there are not contradictory samples in the training set.

[2]The subset is available at https://www.w3.org/TR/wordnet-rdf/.

3.3. Data preparation

Two approaches of dataset preparation were selected. In the first approach, negative examples were generated globally and then the dataset was split into three subsets: training, validation, and test dataset. A proportion 80/10/10 was used for WordNet and OmniScience branches. For the UMLS split 90/5/5 was selected, as the dataset is much smaller and we wanted to give the model a higher number of training examples.

In the second approach, in order to test the ability of the model to generalize to unseen triples, the KGs triples were first divided into three datasets (with the same proportion as before). Then for each of the datasets, negative triples were generated separately. In the end, we excluded the examples that occurred in the intersection of the datasets. A summary of the datasets used for the experiments is presented in Table 4.

Table 2
Number (#) of correct and incorrect triples

Experiment	Performance Experiment			Generalization Experiment		
Dataset	Train Set	Val Set	Test Set	Train Set	Val Set	Test Set
WordNet	49,075 / 99,481	6,096 / 12,474	6,070 / 12,499	48,797 / 148,294	6,210 / 18,659	6,221 / 18,670
UMLS	447/957	28/51	25/53	362/840	47/134	50/127
Physics Branch	4,340 / 12,857	565 / 1,585	499 / 1,651	4,321 / 12,835	541 / 1,613	540 / 1,609
Mathematics Branch	2,221 / 6,536	272/823	277/818	2,215 / 6,561	277/822	277/825

3.4. KG-BERT

KG-BERT is a state-of-the-art method [6] for triple classification tasks. The main idea behind it is to represent the KG triples as text, using their labels to create a lexicalization in natural language and gather contextual sentences from a corpus. This text can then be used to fine-tune the existing pre-trained BERT embeddings, for a classification task. As part of this lexicalization, we explored a set of equivalent ways to represent the notion of hyponymy. In our cases, where we had access to a large corpus *is a concept of* gave the best performance for the model for OmniScience and *is a* for UMLS.

The format of the model's input is as follows: each of the triple elements is separated by the [SEP] token, and at the beginning of the input a [CLS] token is added. Each entity name, as well as the relation's textual description is tokenized. An example of such a representation for the text *"Linear Algebra is a concept of Mathematics"* is: *[CLS] linear algebra [SEP] is a concept of [SEP] mathematics [SEP]*. In the case of the absence of words in the vocabulary, the model is considering their sub-words and is adding specified tokens for the missing parts of the sentence. Cross-entropy is used as a loss function. For our research, we used the code[3] provided by the authors [6] as a basis. We obtained good performance with "bert-base-uncased" BERT base.

[3] https://github.com/yao8839836/kg-bert

4. Experiments and Results

In this section, the performance of the KG triple classification task and its ability to generalize are discussed. By Pp and Pn we refer to the value of precision for positive classes and negative classes respectively, by Rp and Rn we refer to the value of recall for positive and negative classes respectively. Our primary goal is to use the method for KG maintenance, therefore precision for positive examples and recall for negative examples is important (rather than a combined F1 value). With a high score of precision for positive examples, we want to be sure that all returned positives are true positives. With recall for negative examples, we want to be sure that all of the negative examples will be returned by the method, meaning every potential incorrect triple will be returned. Therefore we will focus on these two metrics.

Table 3 presents performance scores for the experiments. WordNet and UMLS were trained using *is a* as relation phrase, and OmniScience branches were trained with *is a concept of*. All of the models were trained using random seed equal to 42, and *bert-base-uncased* as a BERT base. Below we comment in detail on the results.

Table 3
Performance of the model

Dataset	Accuracy	Precision: P/N		Recall: P/N		TP	FN	FP	TN
WordNet	90.33%	84.65%	93.16%	86.03%	92.42%	5,222	848	947	11,552
UMLS	91.02%	80.00%	97.92%	96.00%	88.68%	24	1	6	47
Physics	92.41%	81.23%	96.15%	87.58%	93.88%	437	62	101	1,550
Mathematics	88.12%	75.43%	92.68%	78.70%	91.32%	218	59	71	747
Results generalization:									
WordNet	90.22%	81.82%	92.86%	78.27%	94.20%	4,869	1,352	1,082	17,588
UMLS	81.35%	60.49%	98.96%	98.00%	74.80%	49	1	32	95
Physics	91.53%	80.07%	95.53%	87.04%	93.04%	470	70	112	1,497
Mathematics	91.01%	87.08%	92.11%	75.45%	96.24%	209	68	31	794
Results domain generalization:									
Physics	79.77%	55.1%	89.95%	69.34%	82.92%	346	153	282	1,369
Mathematics	76.98%	52.91%	92.49%	81.95%	75.31%	227	50	202	616

4.1. Performance KG-BERT for different hierarchical Knowledge Graphs

The KG-BERT method applied as a triple classifier is tested using dataset splits described in Table 4. First, the best set of hyperparameters is selected. For this, the model was tested with a combination of: different numbers of epochs, of learning rates, of maximum sentences length, and of training batch size using a grid search approach. We chose the model with the highest accuracy score on the validation set.

We note a high (>88%) accuracy for all of the KGs, and also high scores for all the metrics selected by us as important (Pp and Rn). OmniScience's Physics Branch evaluation gets the highest scores, while the Mathematics Branch gets the lowest for that experiment.

4.2. Model generalization to fully unseen triples

We investigated the model's ability for generalizing on unknown data. We performed two experiments: the first of them (generalization), applied to all four datasets, uses the second split of data reported in Table 4, and the second sampling approach described in subsection 3.3; the second experiment (domain generalization) tested specifically whether a model trained on one OmniScience branch could be applied to another branch. We tested the model trained on one branch with the test set generated for the other branch, for both branches.

Again, we note a high (>90%) accuracy in the generalization experiments for all of the KGs, except from UMLS dataset, where the accuracy is equal to 80%. We discuss the performance on UMLS further in the Discussion section. The Physics Branch classification result gets the highest accuracy score, but the Mathematics Branch achieved the highest scores in independent values for Pp and Rn.

The results for the domain generalization experiment that tested how a model trained on the Mathematics branch of OmniScience performed when classifying examples from the Physics branch (and another way round) are as follows: 80% accuracy for Physics branch being a test set, and 77% for Mathematics branch. Pp is equal to 55% and 53% accordingly, and Rn is equal to 83% and 75%. Even if the accuracy score and Rn are relatively high, we note the low ability of the model to properly classify positive examples (Pp equal to 53-55% in both cases). This means that a model should be trained per domain for a better performance.

4.3. Prediction of long-distance hierarchy using the model

We carried out an experiment to check whether the model can predict a hierarchical relationship between concepts further apart in the hierarchical structure. We created some datasets with triples containing concepts at different levels or hierarchical depth (different hop-levels). As a point of terminology, given the two triples *Optics is a concept of Physics* and *Fiber Optics is a concept of Optics*, we consider the triple *Fiber Optics is a concept of Physics* a 2-hop triple. The three datasets are described in table 4 .

Table 4
Number (#) of correct and incorrect triples for long-distance hierarchy prediction experiment

Dataset	Train Set	Val Set	Test Set	
			1-hop	2-hop
WordNet	49,075 / 99,481	6,096 / 12,474	6,070 / 12,499	96,236 / 288,246
Omni. Physics Branch	4,340 / 12,857	565 / 1,585	499 / 1,651	3,532 / 10,325
Omni. Math Branch	2,221 / 6,536	272 / 823	277 / 818	1,099 / 3,054

For all of the datasets, scores such as accuracy, Pp, and Rn are decreasing with the increase of hop distance. For Pp the decrease is substantial in every dataset (20-30% decrease). For WordNet Pp decreased to 51% for the 2-hop triples, for Physics to 68%, and for Mathematics Branch Pp decreased to 45%. Rn decreased less suddenly (5-8%). For 3-hop, 4-hop, and pairs with more

concepts in-between Pp and Rn continued to decrease. This shows that our model performs well to predict a direct hypernym relationship, but not for hypernym at more than one hop distance.

4.4. Evaluation of classification error output

We performed an error analysis on the incorrectly classified triples output. We found triples such as *blue is a clothing*, *cold is an apartment* or *smoker is a passenger car* were classified as a FP examples in WordNet: these are clearly pure errors of the model's output. In the FP examples for Mathematics and Physics branch of OmniScience, we noticed words overlap between two concepts from the considered triples. For Mathematics branch 43.5% of FP have 3-letter overlapping subwords, almost 30% FP have 4-letter, 23% have 5-letter, and 17.4% have 6-letter overlap. For Physics branch we noted 3-letter overlap for 36.63% of FP, 28% FP noted 4-letter, 21.8% have a 5-letter, and almost 20% have 6-letter overlap. Therefore, we can see that the model has difficulties with establishing the hierarchy between concepts with a large lexical overlap in their naming.

5. Discussion

Performance scores across the selected metrics for QA are high (Precision for positive triples and Recall for negative triples). For every dataset, we noted an accuracy score above 88%. Accuracy for the generalization experiment, where we wanted to see how the model can deal with examples that it did not use for training, yielded decent results (accuracy above 80%). However, testing a model build on one OmniScience branch on another scientific domain of the same KG did not give accurate results. Therefore, the model can generalize well, but needs to be trained per domain.

The reproducibility of the method tested here is a strong point: we showed that this approach can be used on four very different datasets used in different research contexts. However, the accuracy for the UMLS dataset was not as high as the others. The main reason for the lack of performance is the small size of the sample. We further investigated the converge of the model, and concluded that for OmniScience's Physics and Mathematics datasets, the model started to learn with around 1.7k examples. Moreover, for the Physics branch we noted good results from 6.8k training examples (40% of the data), and for Mathematics branch from around 5.6k (65 % of the data). These proportions should be taken into consideration when applying our method.

Our recommendation on how to prepare the model for hierarchy evaluation is as follows:

- the model should be trained per domain,
- relation sentence (or lack of it) could be selected as one of the hyperparameters, and used for the models' optimization,
- proper negative sampling strategy should be selected, the one that can be the most representative strategy of the possible errors in the KGs,
- depending on the scientific domain, different BERT bases could be selected (e.g. *bio-clinical-bert* [23] or *sci-bert* [24] could be used for KGs in the medical or biological domain).

6. Conclusion

In this study, we explored the use of contextual word embeddings for quality assessment of knowledge graph hierarchies. We used KG-BERT, which uses textual information about KGs triples to enrich structure-based embeddings. We described how to use this approach for quality assessment and showed that it works well for both large scale corporate knowledge graphs as well as subsets of one publicly available knowledge graph (WordNet). Moreover, we tested whether the model can generalize across unseen examples and between domains.

We see different paths for future work. First, exploring how to apply the method for the KG maintenance in a real-life practical settings: besides using this framework for a global QA pipeline in real life, this method can also be tested to propose a candidate placement in an existing KG for an incoming candidate concept (by assessing the plausibility of all possible combinations of that candidate with existing concept from a given domain or KG). We have observed good empirical results, but still have to test the idea at scale for the automatic development of KGs.

Secondly, testing the method on other KGs, particularly KGs that have a less consistent hierarchical structure would add value to our understanding of the limitations of the use of BERT embeddings for quality assessment.

In terms of QA methods for the triples assessment, a gold set based approach (assessing the quality of a KG by rebuilding it and comparing it with the original set) could help assessing the feasibility of this method for the automatic generation of large KGs.

References

[1] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, J. Taylor, Industry-scale knowledge graphs: Lessons and challenges, Commun. ACM 62 (2019) 36–43. URL: https://doi.org/10.1145/3331166. doi:10.1145/3331166.

[2] V. Malaisé, A. Otten, P. Coupet, Omniscience and extensions – lessons learned from designing a multi-domain, multi-use case knowledge representation system, in: C. Faron Zucker, C. Ghidini, A. Napoli, Y. Toussaint (Eds.), Knowledge Engineering and Knowledge Management, Springer International Publishing, Cham, 2018, pp. 228–242.

[3] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J. E. L. Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A.-C. N. Ngomo, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, 2021. arXiv:2003.02320.

[4] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, Quality assessment for linked data: A survey, Semantic Web 7 (2015) 63–93. doi:10.3233/SW-150175.

[5] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[6] L. Yao, C. Mao, Y. Luo, KG-BERT: BERT for knowledge graph completion, CoRR abs/1909.03193 (2019). URL: http://arxiv.org/abs/1909.03193. arXiv:1909.03193.

[7] R. Wang, D. Strong, Beyond accuracy: What data quality means to data consumers, J. Manag. Inf. Syst. 12 (1996) 5–33.

[8] B. Stvilia, L. Gasser, M. Twidale, L. Smith, A framework for information quality assessment, Journal of the Association for Information Science and Technology 58 (2007) 1720–1733. doi:10.1002/asi.20652.

[9] H. Chen, G. Cao, J. Chen, J. Ding, A Practical Framework for Evaluating the Quality of Knowledge Graph, 2019, pp. 111–122. doi:10.1007/978-981-15-1956-7_10.

[10] J. Raad, C. Cruz, A survey on ontology evaluation methods, 2015. doi:10.5220/0005591001790186.

[11] Y. Dai, S. Wang, N. Xiong, W. Guo, A survey on knowledge graph embedding: Approaches, applications and benchmarks, Electronics 9 (2020) 750. doi:10.3390/electronics9050750.

[12] A. Bordes, N. Usunier, A. García-Durán, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: NIPS, 2013.

[13] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes, in: AAAI, 2014.

[14] H. Lin, Y. Liu, W. Wang, Y. Yue, Z. Lin, Learning entity and relation embeddings for knowledge resolution, Procedia Computer Science 108 (2017) 345–354. URL: https://www.sciencedirect.com/science/article/pii/S1877050917305628. doi:https://doi.org/10.1016/j.procs.2017.05.045, international Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

[15] H. Xiao, M. Huang, X. Zhu, Transg : A generative model for knowledge graph embedding, 2016, pp. 2316–2325. doi:10.18653/v1/P16-1219.

[16] M. Nickel, V. Tresp, H.-P. Kriegel, A three-way model for collective learning on multi-relational data., 2011, pp. 809–816.

[17] B. Yang, W.-t. Yih, X. He, J. Gao, l. Deng, Embedding entities and relations for learning and inference in knowledge bases (2014).

[18] J. J. Odell, Advanced object-oriented analysis and design using UML, Cambridge ; New York : Cambridge University Press ; New York : SIGS Books, 1998. URL: http://www.loc.gov/catdir/toc/cam024/97018368.html, includes bibliographical references and index.

[19] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, Language, Speech, and Communication, MIT Press, Cambridge, MA, 1998.

[20] A. Benitez, S. Chang, Image classification using multimedia knowledge networks, Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429) 3 (2003) III–613.

[21] O. Bodenreider, The unified medical language system (umls): Integrating biomedical terminology, 2004.

[22] B. Athiwaratkun, A. Wilson, Hierarchical density order embeddings, ArXiv abs/1804.09843 (2018).

[23] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. URL: https://www.aclweb.org/anthology/W19-1909. doi:10.18653/v1/W19-1909.

[24] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, in: EMNLP, 2019.