# Scalable Few-Shot Learning of Robust Biomedical Name Representations

**Pieter Fivez**
CLiPS Research Centre
University of Antwerp
pieter.fivez@uantwerpen.be

**Simon Šuster**
Faculty of Engineering and Information Technology
University of Melbourne
simon.suster@unimelb.edu.au

**Walter Daelemans**
CLiPS Research Centre
University of Antwerp
walter.daelemans@uantwerpen.be

## Abstract

Recent research on robust representations of biomedical names has focused on modeling large amounts of fine-grained conceptual distinctions using complex neural encoders. In this paper, we explore the opposite paradigm: training a simple encoder architecture using only small sets of names sampled from high-level biomedical concepts. Our encoder post-processes pretrained representations of biomedical names, and is effective for various types of input representations, both domain-specific or unsupervised. We validate our proposed few-shot learning approach on multiple biomedical relatedness benchmarks, and show that it allows for continual learning, where we accumulate information from various conceptual hierarchies to consistently improve encoder performance. Given these findings, we propose our approach as a low-cost alternative for exploring the impact of conceptual distinctions on robust biomedical name representations. Our code is open-source and available at www.github.com/clips/fewshot-biomedical-names.

## 1 Introduction

Recent research in biomedical NLP has focused on learning robust representations of biomedical names. To achieve robustness, an encoder should represent the semantic similarity and relatedness between different names (e.g. by their closeness in the embedding space), while its embeddings should also remain as transferable and generally applicable as self-supervised pretrained representations.

Prior research into robust representations has shown three distinct tendencies. Firstly, research typically focuses on encoders with complex neural architectures and a large amount of parameters. As compensation for this complexity, such models can be heavily regularized during training, e.g. by tying the output of a nested LSTM to a pooled embedding of its input representations (Phan et al., 2019), or by integrating a finetuned BERT model with sparse lexical representations (Sung et al., 2020).

| Chapter V: Mental and behavioural disorders | |
|---|---|
| **F34** | **F63** |
| Persistent mood disorders | Habit and impulse disorders |
| F34.0 | F63.0 |
| *Cyclothymia* | *Pathological gambling* |
| F34.1 | F63.1 |
| *Dysthymia* | *Pyromania* |

Table 1: Example of how reference names are grouped together within the ICD-10 hierarchy of disorders.

Secondly, encoders are typically trained on fine-grained concepts from biomedical ontologies such as the UMLS, i.e., concepts with no child nodes in the ontological directed graph. Small synonym sets of such fine-grained concepts are readily available as training data, and often serve as evaluation data for normalization tasks to which trained encoders can be applied.

Lastly, as a result of using fine-grained concepts, vast amounts of biomedical names are needed to model the large collection of fine-grained distinctions present in ontologies. For instance, Phan et al. (2019) train their encoder on 156K disorder names. These three tendencies share an underlying assumption: complex neural encoder architectures can learn biomedical semantics by generalizing in a bottom-up fashion from large amounts of fine-grained semantic distinctions, if provided with sufficient quantities of training data.



*Proceedings of the BioNLP 2021 workshop*, pages 23–29

However, it is not self-evident that such an approach is the most effective way to achieve general-purpose biomedical name representations. For instance, it does not directly address what conceptual distinctions are actually *relevant* to improve representations for downstream NLP applications. Finding and exploiting relevant distinctions can be an empirical question, and as such require low-cost exploration of various conceptual hierarchies. Such a heuristic search is expensive in the current paradigm.

In this paper, we explore a scalable few-shot learning approach for robust biomedical name representations which is orthogonal to this paradigm. We investigate to what extent we can fit a simple encoder architecture using only a small selection of data, with a limited amount of concepts containing only a few samples each (i.e., few-shot learning). To this end, we don't use fine-grained concepts for training, but more general higher-level concepts which span a large range of fine-grained concepts. Table 1 gives an example of such a larger grouping of biomedical names.

This paper offers two main contributions. Firstly, our proposed approach offers an alternative for training biomedical name encoders with much lower computational cost, both for training and inference at test time. It is applicable to large-scale hierarchies containing at least ten thousands of names and is equally effective for different types of pretrained representations when tested on various biomedical relatedness benchmarks. Secondly, we show that this approach allows for low-cost continual learning from multiple concept hierarchies, and as such can help with the accumulation of relevant domain-specific information for downstream biomedical NLP tasks.

## 2 Approach

Our approach is similar to supervised post-processing techniques of word embeddings such as retrofitting and counterfitting (Faruqui et al., 2015; Mrkšić et al., 2016), but instead post-processes pretrained representations of biomedical names.

### 2.1 Encoder architecture

Our encoder architecture is a feedforward neural network with Rectified Linear Unit (ReLU) as non-linear activation function. This neural network transforms a pretrained representation of a biomedical name, after which this transformation is averaged with the pretrained representation:

| | **min** | **max** | **mean** | **stdev** |
|---|---|---|---|---|
| ICD-10 | 247 | 40,519 | 3,414 | 8,693 |
| SNOMED-CT | 397 | 19,114 | 3,532 | 4,094 |
| (+ ambiguous | 1,108 | 23,915 | 4,990 | 5,134) |

Table 2: Descriptive statistics about the number of names per concept for our training data.

$$f(n) = \frac{enc(u_n) + u_n}{2} \quad (1)$$

where $f(n)$ is the output representation for a biomedical name, $u_n$ is its pretrained input representation, and $enc$ is the feedforward neural network which transforms the input representation. The averaging step ensures that the encoder architecture learns to update the pretrained input representation rather than create an entirely new representation. This makes our model more robust against overfitting in few-shot learning settings.

### 2.2 Training objectives

Our training objectives are based on the state-of-the-art BNE model by Phan et al. (2019) and the DAN model by Fivez et al. (2021b), which generalizes the BNE model to any hierarchical level of biomedical concepts. Our framework requires a set of concepts $C$, where each concept $c \in C$ contains a set of concept names $C_n$. The set of biomedical names $N$ contains the union of all those sets of concept names. We propose a simple multi-task training regime which applies two training objectives to each biomedical name $n \in N$. We use cosine distance as distance function $d$ for both objectives.

**Semantic similarity** We enforce embedding similarity between names that are from the same concept by using a siamese triplet loss (Chechik et al., 2010). This loss forces the encoding of a biomedical name $f(n)$ to be closer to the encoding of a semantically similar name $f(n_{pos})$ than that of an encoded negative sample name $f(n_{neg})$, within a specified (possibly tuned) margin:

$$\begin{aligned} pos &= d(f(n), f(n_{pos})) \\ neg &= d(f(n), f(n_{neg})) \\ L_{sem} &= max(pos - neg + margin, 0) \end{aligned} \quad (2)$$

To select negative names during training we apply distance-weighted negative sampling (Wu et al.,

2017) over all training names, since this has been proven more effective than hard or random negative sampling.

**Conceptually grounded regularization** To prevent the model from overfitting on the semantic similarity objective, we regularize it by grounding the output representations to a stable and meaningful target. Simple approximations of prototypical concept representations can already be very effective as targets (Fivez et al., 2021a). Following the model by Fivez et al. (2021b), we use a grounding target which is applicable to any level of categorization, from fine-grained concept distinctions to higher-level groupings of names. This target is a compromise between the *contextual meaningfulness* and *conceptual meaningfulness* objectives of the BNE model. Rather than constraining a name encoding either to its pretrained name representation or to a pretrained representation of its concept, we minimize the distance to the average of both pretrained representations:

$$\begin{aligned} u_c &= \frac{1}{|C_n|} \sum_{n \in C_n} u_n \\ u_{ground} &= \frac{u_c + u_n}{2} \\ L_{ground} &= d(f(n), u_{ground}) \end{aligned} \tag{3}$$

where the concept representation $u_c$ is approximated by averaging each pretrained embedding representation $u_n$ from the set of names $C_n$ belonging to the concept.

This constraint implies that the dimensionality of the encoder output should be the same as that of the input. However, if the input dimensionality is smaller than the desired output dimensionality, this could be solved using e.g. random projections, which work well for increasing the dimensionality of neural encoder inputs (Wieting and Kiela, 2019).

**Multi-task loss** Our multi-task loss sums the losses of the 2 training objectives:

$$L = \alpha L_{sem} + \beta L_{ground} \tag{4}$$

where $\alpha$ and $\beta$ are possible weights for the individual losses. Since both losses directly reflect cosine distances, they are similarly scaled and don't require weighting to work properly. In our experiments, $\alpha = \beta = 1$ showed the most robust performance along all settings.

### 2.3 Training data

We extract sets of high-level concepts and their constituent names from 2 large-scale hierarchies of disorder concepts, ICD-10 and SNOMED-CT. Table 2 gives an overview of our data distributions.

**ICD-10** We use the 2018 version of the ICD-10 coding system.[1] We select the 21 chapters as concept labels, and assign the reference name of each code in a chapter to its concept label. Table 1 gives an example of how such a grouping includes diverse semantic relations.

**SNOMED-CT** We use the 2018AB release of the UMLS ontology[2] to extract a directed ontological graph of SNOMED-CT concepts. We then select the first-degree child nodes of concept *C0012634*, which is the parent concept for all disorders. We then remove those children which are direct parents to other selected children, since they are redundant for our purpose.

This leaves us with 87 concepts, to which we assign the reference terms of all their child concepts in the ontological graph as biomedical names. To make this setup directly comparable to our ICD-10 setup, we select the 21 largest concepts. Finally, we leave out ambiguous names which belong to multiple concepts. Table 2 shows the impact on the data distribution.

## 3 Experiments and discussion

### 3.1 Pretrained representations

We experiment with 3 pretrained name representations. As a first baseline, we use 300-dimensional **fastText** (Bojanowski et al., 2017) word embeddings which we train on 76M sentences of preprocessed MEDLINE articles released by Hakala et al. (2016). We use average pooling (Shen et al., 2018) to extract a 300-dimensional name representation. As a second baseline, we average the 728-dimensional context-specific token activations of a name extracted from the publicly released **BioBERT** model (Lee et al., 2019).

As state-of-the-art reference, we extract 200-dimensional name representations using the publicly released pretrained **BNE** model with skipgram word embeddings, BNE + $SG_w$,[3] which was trained on approximately 16K synonym sets of disease

[1] https://www.cdc.gov/nchs/icd
[2] https://uts.nlm.nih.gov/home.html
[3] https://github.com/minhcp/BNE

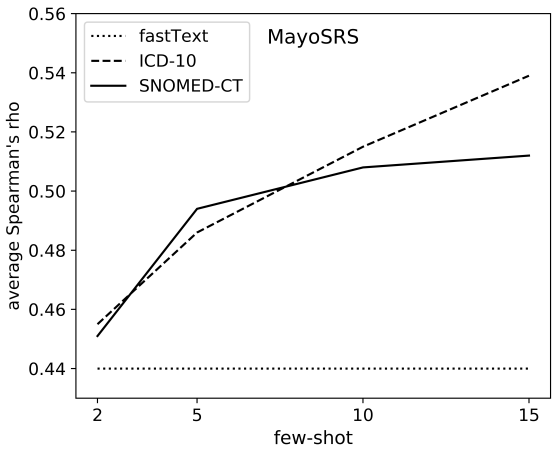


Figure 1: Few-shot performance for fastText encoders on MayoSRS, averaged over 5 random samples.

concepts in the UMLS, containing 156K disease names.

### 3.2 Training details

We randomly sample a small fixed amount of names from each concept in our training data as actual few-shot training names. We then randomly sample the same amount of names as validation data to calculate the multi-task loss as stopping criterion. This criterion is also used to finetune the size of the encoder network. Using only 1 hidden layer proved best in all settings, which leaves only the dimensionality of this layer to be tuned.

Our encoder network is implemented in PyTorch (Paszke et al., 2019). Adam optimization (Kingma and Ba, 2015) is performed on a batch size of 16, using a learning rate of 0.001 and a dropout rate of 0.5. Input strings are first tokenized using the Pattern tokenizer (Smedt and Daelemans, 2012) and then lowercased. We use a triplet margin of 0.1 for the siamese triplet loss $L_{sem}$ defined in Equation 2.

### 3.3 Results

We evaluate our trained encoders on 3 biomedical benchmarks of semantic relatedness and similarity, which allow to compare similarity scores between name embeddings with human judgments of relatedness. MayoSRS (Pakhomov et al., 2011) contains multi-word name pairs of related but different fine-grained concepts. UMNSRS (Pakhomov et al., 2016) contains only single-word pairs, and makes a distinction between *relatedness* and *similarity*, which is a more narrow form of relatedness. Finally, EHR-RelB (Schulz et al., 2020) is much larger than the other benchmarks, and contains multi-word concept pairs which are chosen based on co-occurrence in electronic health records. This ensures that the evaluated concept pairs are actually relevant in function of downstream applications such as information retrieval.

| | EHR-RelB (rel) | MayoSRS (rel) | UMNSRS (rel) | UMNSRS (sim) |
|---|---|---|---|---|
| BioSyn | 0.45 | 0.50 | 0.40 | 0.42 |
| Fivez et al. (2021a) | | **0.67** | **0.56** | 0.56 |
| fastText | 0.39 | 0.44 | 0.47 | 0.48 |
| BioBERT | 0.34 | 0.23 | 0.18 | 0.26 |
| BNE | 0.47 | 0.63 | 0.54 | 0.58 |
| **SNOMED** | | | | |
| fastText | 0.43 | 0.51 | 0.46 | 0.51 |
| BioBERT | 0.40 | 0.31 | 0.32 | 0.38 |
| BNE | 0.53 | 0.63 | 0.55 | **0.60** |
| **ICD-10** | | | | |
| fastText | 0.43 | 0.55 | 0.52 | 0.56 |
| BioBERT | 0.35 | 0.34 | 0.32 | 0.38 |
| BNE | 0.51 | 0.65 | **0.56** | **0.60** |
| **S → I** | | | | |
| fastText | 0.44 | 0.55 | 0.46 | 0.52 |
| BioBERT | 0.39 | 0.33 | 0.35 | 0.42 |
| BNE | **0.54** | **0.67** | 0.52 | 0.58 |
| **I → S** | | | | |
| fastText | 0.45 | 0.54 | 0.46 | 0.51 |
| BioBERT | 0.39 | 0.33 | 0.37 | 0.42 |
| BNE | **0.54** | **0.67** | 0.53 | 0.58 |

Table 3: Spearman's rank correlation coefficient between human judgments and similarity scores of name embeddings, reported on semantic similarity (sim) and relatedness (rel) benchmarks. The highest score is denoted in bold; the second highest is underlined.

We average all test results over 5 different random training samples. We use cosine similarity as similarity score for all baseline representations and trained encoders. Figure 1 shows the impact of the amount of few-shot training names on performance when using fastText representations. Our model already substantially improves over the baseline with only 5 names per concept (105 in total), and maintains consistent improvement up to 15 few-shot names. This confirms that our approach is well-suited to anticipate expected improvements from training on large-scale hierarchies.

Table 3 shows the results on all benchmarks for 15-shot learning. All encoders were tuned to 9,600 hidden dimensions. We include two state-of-the-art biomedical name encoders in our comparison. Firstly, BioSyn (Sung et al., 2020) sums the weighted inner products of fine-tuned BioBERT representations and sparse TF-IDF representations into one similarity score between two names. The pre-trained model[4] for which we report results was

[4] https://github.com/dmis-lab/BioSyn

| | | |
|---|---|---|
| Parent concept | C0042075 | |
| Parent concept name | *disorder of the urinary system* | |
| Validation mention | **urinary hesitancy** | |
| Top 10 ranking | **15-shot BNE** | **BNE** |
| | nebulous urine | nebulous urine |
| | calculus of lower urinary tract ( disorder ) | calculus of lower urinary tract ( disorder ) |
| | urinary obstruction due to nodular prostate ( disorder ) | urinary obstruction due to nodular prostate ( disorder ) |
| | double kidney and/or pelvis | double kidney and/or pelvis |
| | covered exstrophy of bladder ( disorder ) | genital oedema |
| | nephropathy caused by aminoglycoside ( disorder ) | perineal laceration during delivery , nos |
| | renal vein thrombosis | abdominal hernia |
| | benign tumour of urethra | covered exstrophy of bladder ( disorder ) |
| | injury of male urethra | heart :[ weak ] or [ failure nos ] ( disorder ) |
| | postprocedural bulbous urethral stricture | hourglass contraction of uterus |

Table 4: A comparison between the rankings of 315 SNOMED-CT training names for the validation mention *urinary hesitancy*. Non-matching names are underlined. While the pretrained BNE model makes various topical associations, our 15-shot model using the BNE representations as input has learned to cluster around the semantics of urinary tract disorders.

trained on the NCBI disease benchmark (Doğan et al., 2014) for biomedical entity normalization. Secondly, we include the results of the conceptually grounded Deep Averaging Network by Fivez et al. (2021a), which was trained on SNOMED-CT synonym sets mapped into larger ICD-10 categories.

The results show various trends. Firstly, almost all trained encoders improve over their input baselines for all benchmarks, regardless of the type of input representation. Secondly, the performance increase is consistent for both ICD-10 and SNOMED-CT, even as their conceptual hierarchies are substantially different. Lastly, we also look at continual learning from SNOMED-CT to ICD-10 (**S** $\rightarrow$ **I**) or vice versa (**I** $\rightarrow$ **S**), where we use the output of the first model as input representations to train the second model. This approach leads to systematic improvements for all representation types, including the state-of-the-art BNE representations. In other words, we provide tangible empirical evidence that few-shot robust representations can allow for continual specialization in biomedical semantics.

To better understand how our few-shot learning approach can have a visible impact on various relatedness benchmarks, Table 4 gives an example of nearest neighbor names from the training set of SNOMED-CT names for the validation mention *urinary hesitancy*. While the pretrained BNE model makes various topical associations, our 15-shot model using the BNE representations as input has learned to cluster around the semantics of urinary tract disorders. As this already generalizes to validation mentions, we can expect the model to transfer this information to downstream applications wherever urinary tract disorders are relevant. This applies to all 21 high-level topics which were simultaneously encoded for both the ICD-10 and SNOMED-CT ontologies.

## 4 Conclusion and future work

We have proposed a novel approach for scalable few-shot learning of robust biomedical name representations, which trains a simple encoder architecture using only small subsamples of names from higher-level concepts of large-scale hierarchies. Our model works for various pretrained input embeddings, including already specialized name representations, and can accumulate information over various hierarchies to systematically improve performance on biomedical relatedness benchmarks. Future work will investigate whether such improvements trickle down properly to downstream biomedical NLP tasks.

## Acknowledgments

We would like to thank the anonymous reviewers for their feedback. This research was carried out in the framework of the Accumulate VLAIO SBO project, funded by the government agency Flanders Innovation & Entrepreneurship (VLAIO). This research also received funding from the Flemish Government (AI Research Program).

## References


Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. 2010. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11:1109–1135.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics.

Pieter Fivez, Simon Suster, and Walter Daelemans. 2021a. Conceptual grounding constraints for truly robust biomedical name representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2440–2450, Online. Association for Computational Linguistics.

Pieter Fivez, Simon Suster, and Walter Daelemans. 2021b. Integrating higher-level semantics into robust biomedical name representations. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 49–58, online. Association for Computational Linguistics.

Kai Hakala, Suwisa Kaewphan, Tapio Salakoski, and Filip Ginter. 2016. Syntactic analyses and named entity recognition for PubMed and PubMed Central — up-to-the-minute. *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 102–107.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations (ICLR)*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, California. Association for Computational Linguistics.

Serguei V.S. Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B. Melton. 2016. Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*, 32(23):3635–3644.

Serguei V.S. Pakhomov, Ted Pedersen, Bridget McInnes, Genevieve B. Melton, Alexander Ruggieri, and Christopher G. Chute. 2011. Towards a framework for developing semantic relatedness reference standards. *Journal of Biomedical Informatics*, 44:251–265.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Minh C. Phan, Aixin Sun, and Yi Tay. 2019. Robust representation learning of biomedical names. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3275–3285, Florence, Italy. Association for Computational Linguistics.

Claudia Schulz, Josh Levy-Kramer, Camille Van Assel, Miklos Kepes, and Nils Hammerla. 2020. Biomedical concept relatedness – a large EHR-based benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6565–6575, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline needs more love: On simple word-embedding based models and associated pooling mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 440–450.

Tom De Smedt and Walter Daelemans. 2012. Pattern for Python. *Journal of Machine Learning Research*, 13:2031–2035.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

John Wieting and Douwe Kiela. 2019. No training required: Exploring random encoders for sentence classification. In *International Conference on Learning Representations*.

Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krahenbuhl. 2017. Sampling matters in deep embedding learning. In *ICCV*.