Integrating Higher-Level Semantics into 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

Neural encoders of biomedical names are typically considered robust if representations can be effectively exploited for various downstream NLP tasks. To achieve this, encoders need to model domain-specific biomedical semantics while rivaling the universal applicability of pretrained self-supervised representations. Previous work on robust representations has focused on learning low-level distinctions between names of fine-grained biomedical concepts. These fine-grained concepts can also be clustered together to reflect higher-level, more general semantic distinctions, such as grouping the names *nettle sting* and *tick-borne fever* together under the description *puncture wound of skin*. It has not yet been empirically confirmed that training biomedical name encoders on fine-grained distinctions automatically leads to bottom-up encoding of such higher-level semantics. In this paper, we show that this bottom-up effect exists, but that it is still relatively limited. As a solution, we propose a scalable multi-task training regime for biomedical name encoders which can also learn robust representations using only higher-level semantic classes. These representations can generalise both bottom-up as well as top-down among various semantic hierarchies. Moreover, we show how they can be used out-of-the-box for improved unsupervised detection of hypernyms, while retaining robust performance on various semantic relatedness benchmarks. Our code is open-source and can be found at www.github.com/clips/higherlevelsemantics.

1 Introduction

Recent work on representation learning for biomedical names has mainly involved the training of neural encoder architectures such as LSTMs (Kartsaklis et al., 2018) or Transformers (Sung et al., 2020; Kalyan and Sangeetha, 2020) to finetune name representations for biomedical normalization tasks. Such representations are often tailored towards normalization tasks (e.g. linking names to corresponding concept identifiers), without providing explicit guarantees about their transferability to other use contexts and applications. As a solution for this issue, the Biomedical Name Encoder (BNE) model (Phan et al., 2019) has been proposed as a comprehensive framework for robust and transferable representations.

According to this framework, the robustness of biomedical name representations is characterized along three dimensions. Firstly, semantic similarity between names should be reflected by their closeness in the embedding space. Secondly, the variety of textual contexts in which a name appears should be somehow represented in the encoding. Lastly, a name embedding should be sufficiently close to a pretrained prototypical representation of its conceptual meaning, e.g. a representation of its corresponding concept identifier from a biomedical ontology.

Such a multi-task model can be effectively trained using synonym sets extracted from ontologies such as the UMLS or SNOMED-CT. However, these synonym sets typically reflect only fine-grained distinctions between the lowest-level concepts from ontologies. If robust name representations should truly reflect semantic similarity in general, then the assumption is being made that training on such fine-grained synonym sets learns biomedical semantics in a bottom-up way, expecting names of lower-level concepts to spontaneously form relevant higher-level clusters.

Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, pages 49–58
April 19, 2021. ©2021 Association for Computational Linguistics

Level 1	**C0564444** *wound of skin*			
Level 2	**C0561369** *puncture wound of skin*			
Level 3	**C0561546** *bite wound*		**C0576723** *sting of skin*	
Level 4	**C1302713** *animal bite wound*	**C0275134** *poisoning due to lizard venom*	**C0576722** *animal sting*	**C0576724** *plant sting*
Example name	**tick-borne fever**	**poisoning caused by gila monster venom**	**poisoning by bombus**	**nettle sting**

Table 1: Examples of how names from the SNOMED-CT ontology can be grouped into larger classes using parent concepts in the ontological graph. This allows us to investigate higher-level semantic relations, such as grouping *poisoning by bombus* and *nettle sting* under the concept of *sting of skin*, or e.g. grouping them together with *tick-borne fever* under *puncture wound of skin*.

However, such assumptions have not yet been empirically validated, for instance by showing that an encoder not only learns the differences between names such as *nettle sting* and *tick-borne fever*, but also simultaneously learns that they can be grouped together under the more general description *puncture wound of skin*. Moreover, research on representation learning and hierarchical classification for e.g. computer vision has indicated that neural models can leverage substantially different discriminative information for higher, more general levels of categorization than for more fine-grained lower levels (Hase et al., 2019). Such hierarchical differences can be exploited to generalize from higher to lower levels (Guo et al., 2017; Taherkhani et al., 2019), but they can also be difficult to integrate consistently into a single neural model (Wu et al., 2019).

In this paper, we investigate to what extent robust biomedical name representations can encode higher-level semantics while retaining relevant lower-level fine-grained information as well. To address this research question, we group synonym sets under increasingly coarse-grained semantic categories, using parent-child relations in the ontological graph. Table 1 gives an example of how names from the SNOMED-CT ontology can be grouped into larger classes. Such a hierarchy can be used to train and test a variety of semantic relations between names. For instance, a model might be able to encode that the names *poisoning by bombus* and *nettle sting* can be both described as *sting of skin*, but fail to represent their similarity to *poisoning caused by gila monster venom* as a *puncture wound of skin*. We believe that an evaluation of this nature is a crucial step towards achieving truly robust biomedical name representations, since it clearly requires more semantic inference from the encoder than merely resolving synonyms.

Apart from introducing this evaluation to the field of biomedical NLP, we also show that we can effectively adapt the BNE framework (Phan et al., 2019) to be trained using such large higher-level semantic classes. Most importantly, we replace the BiLSTM (Graves and Schmidhuber, 2005) encoder architecture of the BNE model with a lightweight Deep Averaging Network (DAN) (Iyyer et al., 2015). This allows us to easily scale to large amounts of training data, caused by the explosive amount of possible pairwise combinations between semantically similar names as classes grow larger.

Training on higher-level classes involves additional challenges such as handling imbalanced data distributions as well as implicit hierarchical and semantic differences among names grouped under the same class. Our aim is not to tailor the proposed approach to such artefacts. Rather, the main contribution of this paper is to show that our simple modification of the BNE model is generally applicable to a range of coarse-grained biomedical categorizations, without any finetuning apart from the size of the DAN encoder. As of such, it can be used as a low-cost but effective benchmark for future models that are more specialized.

Our experimental results for hierarchical SNOMED-CT data show that our DAN model improves semantic similarity ranking both in a bottom-up as well as top-down manner along various hierarchies. Interestingly, this observation holds even when we train on a few dozens of very broad categories. We also apply extrinsic evaluations to investigate the transferability of our DAN model. Firstly, we validate the robustness of higher-level representations on semantic relatedness benchmarks. Secondly, we perform unsupervised detection of SNOMED-CT hypernym disorder names which were not observed during training. For this task, our DAN model scores substantially better than the publicly released pretrained BNE model, which was trained on a large amount of fine-grained disorder concepts from SNOMED-CT using an elaborate BiLSTM architecture. These results provide tangible evidence that training name representations on large coarse-grained categories can help to encode exploitable higher-level semantics.

2 Related work

While context-dependent self-supervised representations usually outperform other text representations on a variety of BioNLP problems, such as semantic similarity and question answering, there is no single embedding model for biomedical and clinical texts that is consistently superior and thus can serve as a generally suitable bio-encoder (Tawfik and Spruit, 2020). To this date, the BNE model by Phan et al. (2019) is the most prominent attempt at developing a supervised resource for encoding biomedical names. It uses a multi-task training regime in which it combines objectives from different aspects of deep representation learning, such as a contrastive loss (Le-Khac et al., 2020), conceptual grounding (see e.g. (Kartsaklis et al., 2018)), and explicit regularization of the learned representations (e.g. used by Vulić and Mrkšić (2018)). Our modifications to the original BNE model are informed by such literature.

Our application of a Deep Averaging Network (DAN) (Iyyer et al., 2015) is inspired by a recent subfield of NLP research which has emphasized the effectiveness of random encoders (Wieting and Kiela, 2019) and simple pooling mechanisms of word embeddings. The fastText encoder which we use as a baseline and as input for the DAN is an example of a Simple Word-Embedding-based Model (SWEM) with average pooling (Shen et al., 2018).

3 Encoding model

3.1 Encoder architecture

Our encoder is a Deep Averaging Network (DAN) (Iyyer et al., 2015) which extracts a fixed-size representation for an input name *n*:

$$\begin{aligned} u_n &= \frac{1}{|N_t|} \sum_{t \in N_t} u_t \\ f(n) &= enc(u_n) \end{aligned} \tag{1}$$

where N_t is the bag of tokens from a name, u_t is a pretrained word embedding of a token, u_n is a name embedding created by averaging all the pretrained word embeddings of all tokens, and enc is a feedforward neural network with Rectified Linear Unit (ReLU) as non-linear activation function. As pretrained word embeddings we use 300-dimensional fastText (Bojanowski et al., 2017) representations which we train on 76M sentences of preprocessed MEDLINE articles released by Hakala et al. (2016). This fastText model also allows for constructing word embeddings for out-of-vocabulary tokens by composing character n-gram embeddings.

3.2 Training objectives

Our proposed approach is a simple modification of the multi-task training regime of the BNE model. We use cosine distance as distance function d for all three training objectives.

Semantic similarity The *semantic similarity* objective is a generalization from the synonym similarity objective of the BNE model to any level of relevant semantic similarity. To enforce embedding similarity between names that are semantically related, we use 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.

Contextual meaningfulness The *contextual meaningfulness* objective forces the encoding of a biomedical name to be similar to its local contexts. The summary of these local contexts is approximated by taking the pretrained embedding representation u_n of the name:

$$L_{cont} = d(f(n), u_n) \quad (3)$$

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).

Conceptual grounding The *conceptual grounding* objective is a modification of the conceptual meaningfulness objective of the BNE model. The conceptual meaningfulness objective forces the encoding of a biomedical name to be similar to a prototypical representation of its concept. This concept representation is approximated by averaging the pretrained embedding representations of all the names belonging to the concept:

$$u_p = \frac{1}{|C_n|} \sum_{n \in C_n} u_n \quad (4)$$

While converging to this pretrained target is feasible for small synonym sets, such convergence is unnecessary and overfitting for larger classes of names with graded differences in semantic similarity among the class members. To retain the robustness of the encodings, we only want to pull the names in the direction of their pretrained concepts, rather than minimizing their distance entirely. To this end, we simply take the average of the pretrained name representation and the pretrained concept representation:

$$\begin{aligned} v_{ground} &= \frac{u_p + u_n}{2} \\ L_{ground} &= d(f(n), v_{ground}) \end{aligned} \quad (5)$$

Multi-task setup Our multi-task setup sums the losses of the 3 training objectives:

$$L = \alpha L_{sem} + \beta L_{cont} + \gamma L_{ground} \quad (6)$$

where α, β, and γ are possible weights for the individual losses. Since the 3 losses all directly reflect cosine distances, they are similarly scaled and don't require weighting to work properly. In our experiments, $\alpha = \beta = \gamma = 1$ showed the most robust performance along all settings.

4 Data and task setup

4.1 Extracting hierarchical data

Following previous research (Kotitsas et al., 2019; Camacho-Collados et al., 2018), we use IS-A relations between concepts from the SNOMED-CT[1] ontology as biomedical hypo-hypernymy relations. For direct comparison with the publicly released BNE embeddings, which were trained on all disorder concepts of SNOMED-CT, we use the 2018AB release of the UMLS[2] to extract only those SNOMED-CT concepts which are included in the semantic group of disorders[3], and extract their reference terms as disorder names. While the resulting directed graph should be acyclic, there are many inconsistencies, which we resolve by removing all cyclic edges, similar to the naive approach used by Mougin and Bodenreider (2005).

For our experiments, we select 3 different (yet slightly overlapping) subgraphs of IS-A relations by sampling 3 high-level concepts which have around 10K child concepts in our cleaned graph. We extract consistent taxonomies from these subgraphs by removing relations which form shortcuts between otherwise non-consecutive levels of the taxonomy, and by leaving out dead-end concepts which don't have a path to the required level of specification down the taxonomy. Child concepts can have mutually inclusive relations to multiple higher-level concepts on the same level of categorization.

[1] https://www.snomed.org

[2] https://uts.nlm.nih.gov/home.html

[3] https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml

C1290864	min	max	mean	stdev
Level 1	1	10203	1015	2053
Level 2	1	10203	291	1101
Level 3	1	3840	118	411
Level 4	1	2607	48	195

Table 2: Descriptive statistics about the number of names per class for the different levels sampled from the subgraph with parent concept C1290864 (*disorder of abdomen*). These statistics show that lower levels have less extreme imbalances between classes.

4.2 Data setup

For each subgraph, we select 4 consecutive levels of parent concepts (level 1 is highest, level 4 is lowest). The concepts on these 4 levels are used as class labels for the names from all concepts below level 4. In other words, names belonging to the parent concepts themselves are not used during training: the parent concepts are only used as reference to cluster the names from the lower levels. Table 1 visualizes an example of this process.

This method of aggregating names can lead to very imbalanced classes. Table 2 shows how large this imbalance can get as we go up the hierarchy. While the training regime of our proposed model should be robust against such data artefacts, we want to take a representative test sample across all classes to empirically validate our approach. Therefore, for multiple iterations, we sample one held-out test name for each class on level 4. This test name is then also used for levels 1-3. Afterwards, we carry out the same procedure to sample validation data for calculating the stopping criterion during training. Table 3 shows the distributions of concepts and names used during training, validation, and testing.

4.3 Task setup

We perform 2 tasks on the held-out SNOMED-CT test data to validate our approach. Evidently, we always evaluate on individual levels of categorization. As intrinsic evaluation, we evaluate trained encoders on semantic similarity ranking. We also include the task of unsupervised hypernym detection as extrinsic evaluation. As we don't use the names of higher-level concepts during training, we can exploit them as previously unobserved hypernymic data to show how much higher-level semantics are being modeled by encoders. If the encoder has learned to represent biomedical semantics more effectively, then the name embedding space can reflect that by being more suited for unsupervised detection of hypernyms.

Table 1 gives examples of hypernym names on all 4 levels. Successful hypernym detection for this data implies e.g. that we rank the previously unobserved hypernym *bite wound* over another previously unobserved hypernym *sting of skin* for the name *tick-borne fever*. This task clearly requires more semantic inference than merely resolving synonyms. In this case, the encoder has to represent that ticks are insects that bite instead of sting.

Semantic similarity ranking We evaluate encoders on the ability to reflect semantic similarity between names by their cosine similarity. Given a mention *m* of a biomedical name which belongs to the higher-level class *c*, we have to rank the set of all training names *S* which includes $C_n \subset S$, a set of training names which belong to the same class *c* as the test mention. To rank the biomedical names according to their similarity to the mention, we first encode both the mention *m* as well as every name $n \in S$, and then rank every name n using the cosine similarity between the encoded mention $f(m)$ and the encoded name $f(n)$. We then calculate the Mean Average Precision (mAP) over all test mentions for retrieving training names from the same higher-level class.

Unsupervised hypernym detection Given a test mention *m* of a biomedical name which belongs to the higher-level class *c*, we have to rank the set of all hypernym names *H* belonging to a specific level of categorization. This set includes $C_h \subset H$, the set of hypernym names which belong to the same class *c* as the test mention. To rank the biomedical names according to their similarity to the mention, we first encode both the mention *m* as well as every hypernym name $h \in H$, and then rank every hypernym name h using the cosine similarity between the encoded mention $f(m)$ and the encoded hypernym $f(h)$. We then calculate the Mean Reciprocal Rank (MRR) over all test mentions for retrieving hypernym names from the same higher-level class.

5 Experiments and results

5.1 Reference model and baselines

We compare our DAN model against the the publicly released **pretrained BNE** model with skip-gram word embeddings, BNE + SG$_{\mathrm{w}}$,[4] which was

[4] https://github.com/minhcp/BNE

	C1290864 *disorder of abdomen*	**C0560169** *osteoarthropathy*	**C0263661** *dermatological finding*
Level 1	27	30	35
Level 2	98	86	80
Level 3	248	236	231
Level 4	610	536	602
Lower-level names	24737 / 1557 / 763	20574 / 1335 / 649	25659 / 1567 / 814

Table 3: An overview of the distribution of higher-level classes for the 3 subgraphs used in our experiments. The lower-level names are divided into train / test / validation.

trained on approximately 16K synonym sets of disease concepts in the UMLS, containing 156K disease names. We also include 2 baselines: our 300-dimensional **fastText** name embeddings (defined in Equation 1 in Section 3.1), and averaged 728-dimensional context-specific token activations extracted from the publicly released **BioBERT** model (Lee et al., 2019).

5.2 Training and implementation details

The DAN model is implemented in PyTorch (Paszke et al., 2019). Both the input and output dimensionality are 300 (which is the dimensionality of the input fastText embeddings described in Section 3.1). All encoders for which we report results are finetuned to one hidden layer, which has 76,800 dimensions. Adam optimization (Kingma and Ba, 2015) is performed on a batch size of 64, 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.

To train the model, we iterate over all names in the training data and apply the 3 training objectives for each name in a batch. To avoid overfitting on the largest classes, we always sample one siamese triplet per name, using random sampling for the positive name and distance-weighted sampling for the negative name. As stopping criterion we use the mAP of semantic similarity ranking (as defined in Section 4.3) for held-out validation names: we stop training once this score hasn't improved anymore over 10 epochs. This relaxed stopping criterion allows the model to optimize the subsampled siamese triplet loss in a balanced stochastic way over many epochs without quitting too early.

5.3 Results and discussion

Semantic similarity ranking Table 4 shows the test performance for semantic similarity ranking. First and foremost, the robustness of the Level 1 DAN models is consistently great for all 3 subgraphs. For instance, in the case of the subgraph C1290864 (*disorder of abdomen*), the DAN is trained on only 27 large classes but outperforms the fastText baseline for the 610 classes on Level 4. Secondly, all DAN models generalize both bottom-up and top-down along the hierarchical levels to the extent that they consistently outperform the fastText baseline by a substantial margin.

Thirdly, the slight superiority of BioBERT over fastText for this task is most pronounced for the lowest levels. As we go up in the hierarchy, the difference grows smaller, which leads us to believe that the improvements are not so much of a semantic nature. Interestingly, the pretrained BNE model is competitive with our DAN models for the lower levels, which are still more coarse-grained than the fine-grained distinctions on which the BNE was trained. However, such a bottom-up effect is lacking for the highest levels of categorization. These observations reinforce the notion that both the size (the BNE was trained on 156K disorder names, our models on 20-25K) and the granularity of the data matter for deep representation learning.

Unsupervised hypernym detection Table 5 shows the test performance for unsupervised hypernym detection. These results clearly show trends which are similar to the semantic similarity ranking. Most remarkably, the bottom-up and bottom-down effects are almost as consistent here: the highest-level DAN still outperforms the baselines for the lowest levels and vice versa. One major difference with the results for semantic similarity ranking is the relatively worse performance from BioBERT

	C1290864				C0560169				C0263661			
	1	2	3	4	1	2	3	4	1	2	3	4
DAN level 1	**0.57**	0.50	0.39	0.43	**0.70**	0.44	0.36	0.37	**0.64**	0.55	0.36	0.36
DAN level 2	0.49	**0.58**	0.46	0.48	0.55	**0.58**	0.44	0.44	0.58	**0.59**	0.40	0.39
DAN level 3	0.43	0.51	**0.56**	0.54	0.51	0.51	**0.52**	0.54	0.52	0.52	**0.51**	0.48
DAN level 4	0.38	0.43	0.47	**0.60**	0.45	0.45	0.48	0.58	0.45	0.44	0.41	**0.54**
fastText	0.26	0.27	0.25	0.33	0.36	0.29	0.28	0.32	0.33	0.30	0.24	0.30
BioBERT	0.27	0.29	0.29	0.39	0.38	0.32	0.31	0.37	0.36	0.33	0.27	0.35
BNE	0.35	0.41	0.42	0.57	0.43	0.41	0.45	**0.59**	0.44	0.44	0.39	0.51

Table 4: Test performance of semantic similarity ranking per level, as measured by mAP. The highest score per level of each subgraph is denoted in bold; the second highest score is underlined.

	C1290864				C0560169				C0263661			
	1	2	3	4	1	2	3	4	1	2	3	4
DAN level 1	**0.60**	0.58	0.59	0.68	**0.48**	0.54	0.52	0.63	**0.52**	0.57	0.55	0.62
DAN level 2	0.52	**0.59**	0.62	0.70	0.45	**0.58**	0.56	0.67	0.50	**0.60**	0.57	0.63
DAN level 3	0.55	0.57	**0.66**	0.73	0.41	0.54	**0.58**	0.70	0.48	0.57	**0.62**	0.67
DAN level 4	0.53	0.52	0.63	**0.74**	0.39	0.53	**0.58**	**0.74**	0.46	0.54	0.59	**0.71**
fastText	0.46	0.44	0.53	0.65	0.34	0.47	0.49	0.63	0.38	0.45	0.50	0.59
BioBERT	0.41	0.41	0.50	0.62	0.28	0.41	0.46	0.58	0.39	0.47	0.48	0.59
BNE	0.43	0.50	0.60	0.71	0.42	0.48	0.57	0.70	0.49	0.49	0.54	0.68

Table 5: Test performance for unsupervised hypernym detection per level, as measured by MRR. The highest score per level of each subgraph is denoted in bold; the second highest score is underlined.

here compared to fastText. This is in line with the findings by Yu et al. (2020), who report that BERT does not yield considerable improvement for hypernymy detection in their experiments. It also puts into perspective to what extent we can expect higher-level semantics to be encoded solely through self-supervised methods.

Table 6 gives an example of hypernym rankings for the test mention *poisoning caused by mexican beaded lizard bite*. By clustering similar names together with other bite wounds during training, the DAN model has learned to recognize the test mention as a bite wound. The BNE has failed to do so.

The effectiveness of our unsupervised method using only cosine similarity contrasts with earlier approaches which explicitly require more than cosine similarity to properly work. For example, Vulić and Mrkšić (2018) use vector norms to encode hierarchical hypernymic relations, while other research into hypernymy even requires other geometric spaces than Euclidean space, such as hyperbolic space (Dhingra et al., 2018). Our results can indicate that cosine similarity in Euclidean space still shows potential for encoding these hierarchical relations given the right training objectives.

5.4 Semantic relatedness benchmarks

We also evaluate our name encoders on two biomedical benchmarks of semantic similarity, which allow to compare cosine similarity 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, which also stem from different fine-grained concepts. This benchmark makes a distinction between *similarity* and *relatedness*.

The correlations in Table 7 show that the majority of our trained encoders remain robust out-of-the box, with a large portion of them outperforming the fastText baseline which they use as input. The highest-level model trained on the C0560169 subgraph (*dermatological finding*) is even competitive with the pretrained BNE, having been trained on only 30 classes. All in all, these results confirm that our proposed model is relatively robust against variable granularity of clustering, and is not overly tailored to the data artefacts of one specific sub-

Subgraph	**C0560169**	
Level	3	
Test mention	**poisoning caused by mexican beaded lizard bite**	
Matching hypernyms	bite wound / bite wound (disorder)	
	DAN Level 1	**BNE**
Top 5 ranking	*bite wound (disorder)* *bite wound* open traumatic dislocation of hip, unspecified open traumatic dislocation of hip, unspecified (disorder) open dislocation of phalanx of foot (disorder)	infestation caused by fly larvae (disorder) fly larva infestation infestation caused by fly larvae infestation by fly larvae (disorder) infestation by fly larvae

Table 6: A comparison between our DAN encoder and the BNE reference model for unsupervised hypernym ranking of the Level 3 test mention *poisoning caused by mexican beaded lizard bite*. The DAN model generalizes from the training data to associate the test mention correctly with bite wounds. In the training process, it seems to have clustered bite wounds together with open dislocations. The BNE model apparently associates lizards with infestations by fly larvae, but fails to recognize that there is a bite wound mentioned in the test mention.

.

graph.

5.5 Discussion

While our empirical results are certainly encouraging, the true robustness of our proposed framework remains an open question. Whereas our proposed DAN model remains robust over entire hierarchies for semantic similarity ranking and unsupervised hypernym detection, its relative performance for the semantic relatedness benchmarks is not entirely predictable from those tasks. One the one hand, this likely has to do with the modest sizes of the benchmarks, for which small to very small margins in performance are not very reliable or indicative.

On the other hand, we also have to consider that our finetuned DAN only contains a single, yet very wide, hidden layer. This implies that the encoder network relies more on what can considered to be an elaborate weighted average than a deep multilayer transformation of the input. While this is not very surprising in the context of transferable representations (and emphasizes the effectiveness of exploiting word embeddings according to their full potential in simple ways, as suggested by Wieting and Kiela (2019)), it still raises the question whether there are straightforward regularization alternatives to the contextual meaningfulness objective which can allow for deep transformations with the DAN.

6 Conclusion and future work

In this paper, we have introduced the challenge of integrating higher-level semantics into robust biomedical name representations. We provide a framework to both train and evaluate encoders for

	MayoSRS (rel)	UMNSRS (rel)	UMNSRS (sim)
fastText	0.44	0.47	0.48
Level 1 C0560169	0.42	0.55	0.54
Level 2 C0560169	0.47	0.51	0.50
Level 3 C0560169	0.50	0.51	0.50
Level 4 C0560169	0.50	0.51	0.50
Level 1 C1290864	0.52	0.42	0.46
Level 2 C1290864	0.55	0.46	0.40
Level 3 C1290864	0.53	0.46	0.50
Level 4 C1290864	0.56	0.45	0.50
Level 1 C0263661	0.46	0.49	0.51
Level 1 C0263661	0.51	0.47	0.50
Level 3 C0263661	0.55	0.50	0.53
Level 4 C0263661	0.52	0.50	0.50
Phan et al. (2019)	**0.63**	**0.58**	**0.61**

Table 7: Spearman's rank correlation coefficient between cosine similarity scores of name embeddings and human judgments, reported on semantic similarity (sim) and relatedness (rel) benchmarks. The highest score is denoted in bold; the second highest is underlined.

this task. Moreover, we have proposed a modification of the Biomedical Name Encoder model which is directly applicable to a variety of coarse-grained categorizations. This modification replaces more complex neural architectures with a lightweight Deep Averaging Network encoder, which is easily scalable to the large amounts of required training data, while remaining sufficiently robust. The only important hyperparameter to tune for this encoder is the size of the Feedforward Neural Network.

Experiments indicate that our proposed framework can even be effective using only around 30 coarse-grained classes. This opens up possibilities for applying our framework to data beyond carefully curated ontologies, for instance in self-supervised or semi-supervised settings. Future work will try to understand and define the limits of applying our framework to such settings.

Acknowledgments

We would like to thank the anonymous reviewers for their feedback. This research was funded by the government agency Flanders Innovation & Entrepreneurship (VLAIO) in the framework of the Accumulate VLAIO SBO project, and by 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.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 task 9: Hypernym discovery. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 712–724, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Bhuwan Dhingra, Christopher Shallue, Mohammad Norouzi, Andrew Dai, and George Dahl. 2018. Embedding text in hyperbolic spaces. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pages 59–69, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Alex Graves and Jurgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Yanming Guo, Yu Liu, Erwin M. Bakker, Yuanhao Guo, and Michael S. Lew. 2017. CNN-RNN: a large-scale hierarchical image classification framework. *Multimedia Tools and Applications*, 77:10251–10271.

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.

Peter Hase, Chaofan Chen, Oscar Li, and Cynthia Rudin. 2019. Interpretable image recognition with hierarchical prototypes. In *The Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP-19)*.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Association for Computational Linguistics*.

Katikapalli Subramanyam Kalyan and Sivanesan Sangeetha. 2020. Medical concept normalization in user-generated texts by learning target concept embeddings. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 18–23, Online. Association for Computational Linguistics.

Dimitri Kartsaklis, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Mapping text to knowledge graph entities using multi-sense LSTMs. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1959–1970.

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

Sotiris Kotitsas, Dimitris Pappas, Ion Androutsopoulos, Ryan McDonald, and Marianna Apidianaki. 2019. Embedding biomedical ontologies by jointly encoding network structure and textual node descriptors. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 298–308, Florence, Italy. Association for Computational Linguistics.

P. H. Le-Khac, G. Healy, and A. F. Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.

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

Fleur Mougin and Olivier Bodenreider. 2005. Approaches to eliminating cycles in the UMLS metathesaurus: Naïve vs. formal. In *AMIA Annual Symposium Proceedings*.

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.

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.

Fariborz Taherkhani, Hadi Kazemi, Ali Dabouei, Jeremy E. Dawson, and Nasser M. Nasrabadi. 2019. A weakly supervised fine label classifier enhanced by coarse supervision. In *ICCV*.

Noha S. Tawfik and Marco R. Spruit. 2020. Evaluating sentence representations for biomedical text: Methods and experimental results. *Journal of Biomedical Informatics*, 104.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1134–1145, New Orleans, Louisiana. 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*.

Cinna Wu, Mark Tygert, and Yann LeCun. 2019. A hierarchical loss and its problems when classifying non-hierarchically. *PLOS ONE*, 14(12):1–17.

Changlong Yu, Jialong Han, Peifeng Wang, Yangqiu Song, Hongming Zhang, Wilfred Ng, and Shuming Shi. 2020. When hearst is not enough: Improving hypernymy detection from corpus with distributional models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6208–6217, Online. Association for Computational Linguistics.