Biomedical Interpretable Entity Representations

Diego Garcia-Olano[1,2] **Yasumasa Onoe**[2] **Ioana Baldini**[1]
Joydeep Ghosh[2] **Byron C. Wallace**[3] **Kush R. Varshney**[1]
[1]IBM Research, [2]University of Texas at Austin, [3]Northeastern University
diegoolano@gmail.com, yasumasa@utexas.edu, ioana@us.ibm.com,
ghosh@ece.utexas.edu, b.wallace@northeastern.edu, krvarshn@us.ibm.com

Abstract

Pre-trained language models induce dense entity representations that offer strong performance on entity-centric NLP tasks, but such representations are not immediately interpretable. This can be a barrier to model uptake in important domains such as biomedicine. There has been recent work on general interpretable representation learning (Onoe and Durrett, 2020), but these domain-agnostic representations do not readily transfer to the important domain of biomedicine. In this paper, we create a new entity type system and training set from a large corpus of biomedical texts by mapping entities to concepts in a medical ontology, and from these to Wikipedia pages whose categories are our types. From this mapping we derive *Biomedical Interpretable Entity Representations* (BIERs), in which dimensions correspond to fine-grained entity types, and values are predicted probabilities that a given entity is of the corresponding type. We propose a novel method that exploits BIER's final sparse and intermediate dense representations to facilitate model and entity type debugging. We show that BIERs achieve strong performance in biomedical tasks including named entity disambiguation and entity label classification, and we provide error analysis to highlight the utility of their interpretability, particularly in low-supervision settings. Finally, we provide our induced 68K biomedical type system, the corresponding 37 million triples of derived data used to train BIER models and our best performing model.

1 Introduction

In modern NLP systems, entities are embedded in the same dense vector space as words using vectors from pre-trained (masked) language models (Devlin et al., 2019) that yield contextualized embeddings of tokens. These representations are used as inputs for downstream models built for particular tasks. One issue with such learned representations is that we do not actually know what information they encode. Recent work has shown that deep pre-trained models implicitly learn factual knowledge about entities (Petroni et al., 2019; Roberts et al., 2020), but the embeddings that they provide do not explicitly maintain representations of this knowledge (i.e., the dimensions in learned representations have no *a priori* semantics); consequently, are not directly interpretable. This has motivated the design of knowledge *probing tasks* to measure a factual knowledge implicit in embeddings (Petroni et al., 2019; Poerner et al., 2019).

Recent work (Onoe and Durrett, 2020) has proposed learning interpretable entity representations using an entity typing model and corresponding fine-grained type system that accepts an entity mention and its context. The output represents a high-dimensional sparse embedding whose values correspond to the model's (independently) predicted probabilities that the entity possesses the respective properties defined by the fine-grained type system.

This past work proposed general domain pre-trained Transformer-based (Vaswani et al., 2017) entity typing models trained on Wikipedia or the ultra-fine entity typing system (Choi et al., 2018), yielding 60k and 10k dimensional embeddings, respectively, which can then be used directly in downstream tasks. Such representations can achieve strong results without learning task specific representations. Thus, in addition to providing interpretability, such representations may be particularly useful for tasks with limited supervision.

Such interpretable entity representations for text can be valuable in domains such as biomedicine, because they afford model transparency which may help with model debugging, or simply to instill confidence in model outputs. For example, if one defines a linear layer on top of entity-type representations, learned coefficients are interpretable as weights assigned to specific entity types. One could debug an incorrect prediction by inspecting the induced representation for potentially erroneous types assigned to it. This sort of insight is particularly important in biomedical NLP, given the potential sensitivity of the tasks in the domain, and the high-level expertise of the 'end-users'.

Motivated by these observations, we extend (Onoe and Durrett, 2020) to learn sparse Biomedical Interpretable Entity Representations (BIERs) in which values encode predicted probabilities of an entity belonging to a type from a fine-grained entity type system. Starting from a corpus of PubMed[1] articles on cancer and drugs as our training data, we induce an entity type system by mapping entities in the articles to their associated UMLS concepts, and then mapping the concepts to Wikipedia pages whose categories we use as our types.

We show that learning a typing model on top of such a system realizes strong performance on a variety of biomedical tasks including named entity disambiguation (NED) and entity label classification using simple cosine similarity or Euclidean distance based methods, and we provide an analysis of the results from an interpretabilty perspective. In addition, we propose a simple technique that facilitates debugging and provides a mechanism by which to improve model performance by exploiting both the proposed sparse interpretable type representations and their internal underlying dense counterparts. Finally, we introduce and release a new medical-centric Wikipedia dataset based on (Rosenthal et al., 2019) for use in the task of biomedical NED.

Our specific contributions[2] are as follows:

- We create (and will release) a biomedical entity typing system comprising Wikipedia Categories from pages mapped to UMLS concepts linked to PubMed article entities and learn a model that produces sparse entity representations in which dimensions are imbued with known semantics. We show that these achieve strong performance on biomedical NED and entity label classification tasks.
- We conduct an interpretability analysis and demonstrate a new debugging method that uses the proposed representation's performance on downstream tasks to gain insights into the entity typing model and system.
- We release a medical literature centric Wikipedia dataset for use in the task of biomedical NED.

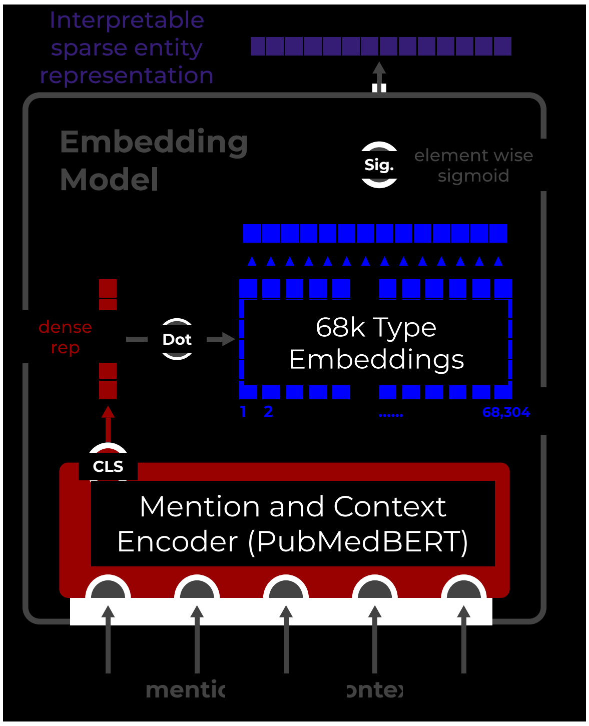


Figure 1: Model architecture from (Onoe and Durrett, 2019) using our 68k biomedical entity type system. A BERT based encoder embeds a mention and context and the output entity embedding contains probabilities for each type.

2 Background: Interpretable Entity Model

We first review the interpretable entity model architecture we extend from (Onoe and Durrett, 2020).

Let $s = (w_1, ..., w_N)$ denote a sequence of input context words, $m = (w_i, ..., w_j)$ denote an entity mention span in s, and $\mathbf{t} \in [0, 1]^{|\mathcal{T}|}$ denote a vector whose values are predicted probabilities corresponding to fine-grained entity types $\mathcal{T}$ from a predefined type system with higher values identifying types most pertaining to m and s .

Given a labeled dataset $\mathcal{D} = \{(m, s, \mathbf{t}^*)^{(1)}, ..., (m, s, \mathbf{t}^*)^{(k)}\}$ the objective is to learn parameters θ of a function f_θ that maps the mention m and its context s to a vector $\mathbf{t}$ that captures salient features of the entity mention within its context. The basic idea is that the

[1] A repository of biomedical literature: http://www.pubmed.gov/.

[2] Code and datasets available at http://github.com/diegoolano/biomedical_interpretable_entity_representations

resultant entity embeddings $\mathbf{t}$ (wherein individual dimensions have explicit semantics) can be used as embeddings in downstream tasks, for example by using basic similarity measures such as dot products or cosine similarities.[3]

The simple model f_θ that produces these embeddings is shown in Figure 1. First, a BERT-based encoder (Devlin et al., 2019) maps inputs m and s to an intermediate dense vector representation. Specifically, the encoder takes as input a token sequence formatted as $\mathbf{x} = \texttt{[CLS]}\ m\ \texttt{[SEP]}\ s\ \texttt{[SEP]}$, where the mention m and context s are segmented into WordPiece tokens (Wu et al., 2016). The hidden vector output corresponding to the `[CLS]` token can be treated as the intermediate dense mention and context representation: $\mathbf{h}_{\texttt{[CLS]}} = \textsc{BertEncoder}(\mathbf{x})$.

A type embedding layer then projects this intermediate representation to a vector whose dimensions correspond to the entity types $\mathcal{T}$ using a single linear layer whose parameters may be viewed as a matrix of type embeddings $\mathbf{E} \in \mathbb{R}^{|\mathcal{T}|\times d}$, where d is the dimension of the mention and context representation $\mathbf{h}_{\texttt{[CLS]}}$. Finally, we apply a sigmoid function to each unnormalized score in the vector to obtain the predicted probabilities that form our entity representation $\mathbf{t}$ (top of Figure 1). We obtain these output probabilities $\mathbf{t}$ by multiplying $\mathbf{E}$ by $\mathbf{h}_{\texttt{[CLS]}}$, followed by an element-wise sigmoid function: $\mathbf{t} = \sigma\left(\mathbf{E}\cdot\mathbf{h}_{\texttt{[CLS]}}\right)$

Following Choi et al. (2018), the training loss we minimize is a sum of binary cross-entropy losses over all entity types $\mathcal{T}$ over all training examples $\mathcal{D}$. That is, we treat each type prediction for each example as an independent binary decision, with shared parameters in the BERT encoder. Our loss $\mathcal{L}$ is:

$$-\sum_i \sum_j t^*_{ij} \cdot \log(t_{ij}) + (1 - t^*_{ij}) \cdot \log(1 - t_{ij}),$$

where i are the data indices, j are indices over types, t_{ij} is the jth component of $\mathbf{t_i}$, and t^*_{ij} is the jth component of $\mathbf{t_i}^*$ that takes the value 1 if the jth type applies to the current entity mention. We fine-tune all parameters in BERT and the type embedding matrix.

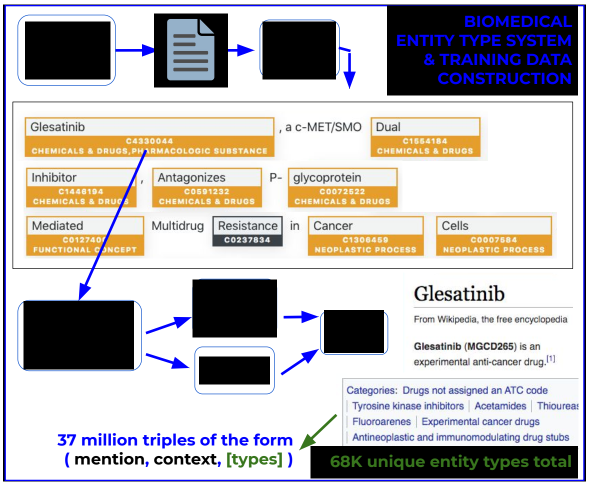

BIOMEDICAL
ENTITY TYPE SYSTEM
& TRAINING DATA
CONSTRUCTION
Glesatinib
C4330044
CHEMICALS & DRUGS,PHARMACOLOGIC SUBSTANCE
, a c-MET/SMO
Dual
C1554184
CHEMICALS & DRUGS
Inhibitor
C1446194
CHEMICALS & DRUGS
,
Antagonizes
C0591232
CHEMICALS & DRUGS
P-
glycoprotein
C0072522
CHEMICALS & DRUGS
Mediated
C0127400
FUNCTIONAL CONCEPT
Multidrug
Resistance
C0237834
in
Cancer
C1306459
NEOPLASTIC PROCESS
Cells
C0007584
NEOPLASTIC PROCESS
Glesatinib
From Wikipedia, the free encyclopedia
Glesatinib (MGCD265) is an experimental anti-cancer drug.[1]
Categories: Drugs not assigned an ATC code
Tyrosine kinase inhibitors | Acetamides | Thioureas
Fluoroarenes | Experimental cancer drugs
Antineoplastic and immunomodulating drug stubs
37 million triples of the form
(mention, context, [types])
68K unique entity types total

Figure 2: Biomedical Entity Type System and dataset construction. Appendix Fig 4 contains example output.

3 Biomedical Interpretable Entity Representations

Biomedical Entity Typing To train an interpretable entity embedding model tailored specifically for biomedical tasks, we must first construct a suitable biomedical entity type system and dataset. PubMed indexes over 30 million biomedical citations across a wide range of topics. To curate a topically focused set of literature, we first used the PubTator tool (Wei et al., 2019) to query PubMed for articles related to drugs used as treatment for cancer; this yielded 461,404 unique citations (titles and abstracts).[4]

We used an off the shelf NER tagger available in `SciSpacy` (Neumann et al., 2019) to identify entity spans within abstracts, and used the `Entity Linker` component to link those entities to concept unique IDs (CUIDs) within the Unified Medical Language System (UMLS) ontology[5].

Next we had to decide on the specific entity type system to use, i.e., the set of labels to attach to entities, and chose Wikipedia as our knowledge base. We used this general knowledge base instead of a specialized ontology (for example, MeSH or SNOWMED CT) primarily because it yielded (many) more diverse entity types per mention, comparatively.

To connect UMLS concepts to Wikpedia pages

[3] Fine-tuning the representations would destroy their interpretability.

[4] We selected the topic of cancer because our work is motivated by a larger project aimed at finding existing evidence that supports repurposing generic drugs for cancer.

[5] UMLS defines around 3 million concepts from a combined 200 source ontologies. Concepts may be identified as having one or more of 127 semantic types which can be used to place them into groupings such as diseases or drugs.

we use the mapping from Cuzzola et al. (2018), which is accurate but incomplete: It provides exact wikipage matches for 221,690 concepts and "close matches" for 26,276 of them, out of a possible 3 million concepts in UMLS. For concepts for which no exact or close match was found, we used SLING (Ringgaard et al., 2017), a framework for frame semantic parsing which allows for querying and resolving wikipages given a search string (in our case, mention surface forms). For high confidence exact or close matches, we return the set of categories found for their combined results. While these results can be slightly noisy, they mostly lead to satisfactory performance.

We filter the entity mentions that compose our final set, as follows. If multiple concept CUIDs are found for a given entity, we include the highest scoring matches within two points of each other provided they all exceed a minimum score threshold of 0.8;[6] Additionally, we only include results that are linked to at least one concept CUID and where an associated Wiki link was mapped to directly via Cuzzola et al. (2018) or via SLING. A schematic of this process is shown in Figure 2. An example working through the entity filtering process is shown in the text of Appendix A. In the end about 12.5% of the mappings from PubMed mentions to Wikipedia categories come via SLING.

After processing, linking and filtering the corpus of PubMed abstracts, we were able to extract 37,357,141 triples of the form (mention, context, [list of categories]). This list of triples contains 68,304 unique categories which we use as the entity type system for training BIERs. Appendix 8 contains a list of the top 100 entity types that appear over these articles and Appendix 5 shows a histogram of entity types per mention. As one contribution, we will release this set of derived triples.

To assess the quality of this dataset, we chose 500 triples at random and asked 4 experts (researchers in biomedicine and ML) to score them on a Likert scale from 1 (low) to 5 (high) for accuracy. Experts assessed how well a PubMed mention from a context sentence maps onto a Wikipedia URL. Average expert scores for the triples were [4.01, 4.13, 4.18, 4.20] (overall mean of 4.13) out of 5. The Fleiss-Kappa score which measures inter-annotator agreement was strong at .69. Additionally 77% of scores are >= 4, and for 93% of the examples at least 3/4 experts agree (73% have unanimous agreement).

BIER entity typing model training and test results We split our derived dataset of biomedical triples into train, validation, and test sets of sizes 31,340,000, 376,071, and 5,641,070, respectively. For comparison, the total data size used by Onoe and Durrett (2020) is 6.1 million and based on the most popular categories of Wikipedia whereas ours only uses categories on pages linked to UMLS.

We trained different BIER models using variants of BERT as an encoder for mentions and contexts. Specifically we considered BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019) and BLURB (Gu et al., 2020) (we will refer to this as PubMedBERT), which constitute the current state of the art for many biomedical tasks.We compute entity typing macro F1 using development examples to check model convergence and use the hyperparameters from Onoe and Durrett (2020).

Debugging BIERs by combining dense and sparse embeddings We propose a technique for debugging using BIER representations that is in part inspired by prior work that used intermediate layer representations of training examples as additional features (Papernot and McDaniel, 2018). Specifically, we propose to debug BIER performance on downstream tasks by examining instances where dense and sparse representations yield different outputs. For each example, BIER models produce an intermediate dense $\mathbf{h}_{\texttt{[CLS]}}$ and interpretable sparse output embeddings t (red and purple, respectively, in Figure 1). We will refer to the two seperate models which use these dense and sparse BIERs embeddings for downstream tasks as `BIERDense` and `BIERSparse` respectively.

After performing inference initially, we gather all test examples where the `BIERDense` makes a correct prediction but `BIERSparse` does not and we place their mention values into a set $\mathcal{Z}$. Additionally, as a diagnostic measure, we consider an 'oracle' approach in which we use the `BIERDense` prediction for all instances in $\mathcal{Z}$, and the `BIERSparse` output otherwise. The intuition is that $\mathcal{Z}$ contains examples for which the intermediate dense embeddings better represent a mention-context than the more interpretable sparse output embeddings from the BIER model.

Because the sparse embeddings are interpretable, this analysis affords fine-grained analysis of which

[6] This is the default threshhold set in `SciSpacy` for concept candidate inclusion.

Dataset	Mentions	Abstracts	Type	Sets
MedMentions	350K	4.3K-PubMed	gold	no
BIERs*	37M	460K-PubMed	silver	no
ClinWikiNED*	10K	35K-Wiki	silver	yes

Table 1: Comparison of BioMed Linked Datasets. BIERs and ClinicalWikiNED datasets described in Section 3 and 4.1 respectivtely

entity types lead to incorrect predictions by the sparse model (but correct predictions using a dense representation). This diagnostic can be used as a benchmark for how well the model could have done had the entity typing model's output better represented the mention-context, or if the model had known to fallback to using the intermediate dense embedding; the former case might be ameliorated via more supervised examples or changes to the type system while the latter could motivate a dynamic approach to making predictions that is a function of model confidence. We show results and analysis using these methods in Section 5.

4 Experimental Setup

To evaluate the utility of the proposed biomedical entity representations, we use them for the tasks of biomedical entity label classification (ELC) and named entity disambiguation (NED). We highlight that these models perform well even without finetuning, which is critical in low- or zero-supervision scenarios.

4.1 NED on Biomedical Wikipedia articles

The NED task connects entity mentions in text with real world entities in a knowledge base by disambiguating the true entity from a list of candidates. We consider the local resolution setting in which each instance features a single entity mention span in the input text and several possible candidates with corresponding descriptions (e.g., the first paragraph of their Wikipedia article).

NED dataset construction While there exist multiple biomedical named entity recognition and linking datasets (Mohan and Li, 2019; Basaldella et al., 2020), we did not find much in the way of publicly available biomedical NED corpora, and we therefore constructed a new dataset, which we will release for use by other researchers. The dataset is based on the set of Wikipages used by Rosenthal et al. (2019), as relevant medical literature which consists of 34,692 medically relevant articles under the 'Clinical Medicine' category [7]. We used SLING[8] to process these articles and were able to retrieve around 1.5 million training examples (mention, context, [categories]) from them.

After obtaining these examples for each entity mention we used the CrossWikis dictionary (Spitkovsky and Chang, 2012) to try to gather between 3 to 5 challenging candidate entities for the example. This range in terms of number of candidates was selected because we wanted to include salient biomedical terms that are difficult to disambiguate; setting a higher number of potential candidates for use with CrossWikis largely gives general and short "popular" candidates (i.e., those that appear often in Wikipedia). This behavior makes sense since many biomedical terms are quite specific and usually only have a few high quality alternative candidates to select from. Additionally, we filter out redirect pages and pages that no longer match the wiki version used to create CrossWikis.

This candidate generation and data content acquisition step filters out considerably the number of available examples. We additionally subsample the dataset to reduce the instances where the "popular" candidate is the correct entity so as to make the task more difficult and to allow for more rare entities to appear in our set. After all the filtering, our ClinicalWikiNED dataset consists of a train/dev/test split of size 5332, 3730, and 800 respectively. Table 1 shows a comparison of the two datasets introduced in this paper with that of one of the largest publicly available linked biomedical datasets(Murty et al., 2018).

Using BIERs for NED Using the BIER architecture, we first train a separate `WikiDescription` model that takes as input a wikipage title as its mention, its first paragraph as the context, and outputs a sparse embedding that predicts the page's categories. As training data, we use any Wikipedia page that contains categories in our biomedical entity type system. We use 2.5 million such (title, descriptions, [categories]) as our training data, and we check for model convergence on a small development set. This model is used so that candidate embedding dimensions will align with our BIER mention-context embeddings.

For each mention m and context s in the test set, we use a BIER model to induce a sparse rep-

[7] https://en.wikipedia.org/wiki/Category:Clinical_medicine

[8] Based on a June 1, 2020 dump of English wikipedia.

Model	Test Acc.	
	Dot Prod	Cosine Sim
BIER-PubMedBERT (ours)	80.1	**84.0**
BIER-SciBERT (ours)	76.4	77.3
BIER-BioBERT (ours)	71.9	75.9
Onoe and Durrett (2020)	63.6	69.8
Popular Prior	73.9	-
PubMedBERT (Gu et al., 2020)	77.6	-
SciBERT (Beltagy et al., 2019)	77.4	-
BioBERT (Lee et al., 2019)	77.9	-

Table 2: BIER zero shot test results vs Logistic Regression Baselines trained on task data for NED task

resentation t. We then go through each candidate c_i for the current test example and use the `WikiDescription` model to retrieve the candidate's sparse output embedding t_{c_i}. Finally, we compute both the cosine similarity and dot product of t with each candidate t_{c_i} and predict the candidate c_i that achieves the highest score for each metric as the true one.

Baseline model for NED We use the EntEval (Chen et al., 2019) framework for our experiments and train a logistic regression classifier using a feature vector composed of the mention-context embedding x_1 and current candidate wiki description embedding x_2 from the set of candidates C_m as a concatenation of x_1, x_2, element-wise product, and absolute difference: $[x_1, x_2, x_1 \odot x_2, |x_1 - x_2|]$. Both x_1 and x_2 are obtained via BERT based models. Training minimizes binary log loss using all negative examples. At test time, inference combines this classifier result with the prior probability of how frequently candidates occur in Wikipedia as follows: $arg\ max_{c \in C_m}[p_{prior}(c) + p_{classifier}(c)]$ to obtain the final candidate prediction. Directly using the most likely prior as predictions yields an accuracy of 73.9%. We emphasize that these baselines are fine-tuned on the task data while the BIER models only do inference on the test set.

Results Table 2 shows the results of the NED experiments. The biomedical BIER model affords improvements over the prior general domain interpretable model (Onoe and Durrett, 2020), showing that the biomedical type system and training is beneficial for this type of task. In addition, the BIER models outperform the baselines without fine-tuning on the training data.

4.2 ELC on Cancer Genetics data

For our entity label classification task we use the Cancer Genetics dataset (Pyysalo et al., 2013) which consists of 10,935 training, 3,634 dev, and 6,955 test examples from 300, 100, and 200 unique PubMed articles, respectively.[9] Given an article title and abstract, mention, and the corresponding entity label, the objective is to predict this label from 16 available coarse labels (see Table 7 in the Appendix for label distribution information).

To assess how well the learned BIER representations fare against comparable baselines, we perform a simple nearest neighbor classification technique using the proposed BIER model variants, the general domain model from Onoe and Durrett (2020), and non-BIER fine-tuned pre-trained language models as standalone encoders.

We first induce dense embeddings for all training examples by passing the mention m and context s through the encoders as `[CLS]` m `[SEP]` s `[SEP]`, and we store the resultant contextualized `[CLS]` embedding $\mathbf{h}_{\texttt{[CLS]}}$ as our dense embedding. For the BIER and Onoe and Durrett (2020) models we also save the final sparse entity embedding $\mathbf{t}$.

We iterate over the test examples and similarly induce dense representations for these $\mathbf{h}^{test}_{\texttt{[CLS]}}$ and (if applicable) sparse representations $\mathbf{t}^{test}$. We find their nearest neighbor (under either $\ell 2$ distance or dot product similarity) from the saved training set of embeddings, and use its label as the prediction. We use the FAISS semantic indexer (Johnson et al., 2017) for storing embeddings and finding nearest neighbors quickly. We are interested in evaluating the off-the-shelf utility of learned representations, and, as such, we do not train or fine-tune the models in any of these cases; rather, training examples are used only for nearest neighbor retrieval.

That said, for completeness we also performed additional experiments in which we do fine-tune models on the task data, with varying amounts of supervision; we are interested especially in low-supervision settings. For the fine-tuning experiment, we add a linear layer on top of the best performing BIER and baseline models, using cross entropy loss as our objective and fine-tuning them for 4 epochs on the training data before performing inference. For the low supervision regime experiment, we show how the best nearest neighbor and

[9] In our experiments we combine the train and dev sets into a single training set.

	Test Acc.			
	L2 Dist		Dot Prod	
Model	Dense	Sparse	Dense	Sparse
BIER-PubMedBERT	85.5	86.8	**88.2**	**87.5**
BIER-SciBERT	70.8	77.0	72.8	76.8
BIER-BioBERT	83.4	85.9	85.6	86.8
Onoe and Durrett (2020)	63.9	55.1	60.0	59.9
PubMedBERT	77.3	-	69.3	-
SciBERT	74.4	-	75.2	-
BioBERT	67.6	-	59.6	-

Table 3: Test accuracy on Cancer Genetics data using a nearest neighbor classifier (k=1) without fine-tuning based on sparse output or intermediate dense embeddings using L2 or Dot Product distance metrics.

fine-tuned models perform when given K examples per class for $K \in [5, 10, 25, 50, 75, 100, 200]$

Results Table 3 shows the results for our first experiment, in which we use untuned representations. We observe that the baseline language model encodings all perform worse than the proposed BIER sparse and dense models, with the exception of SciBERT, which fares better than the sparse BIER model based on SciBERT. Additionally, we see that BERT and Onoe and Durrett (2020) (which is based on BERT) both perform poorly in this biomedical task compared to the other baselines.

Importantly, we notice that the sparse interpretable embedding results for our top performing models (both BIER-PubMedBERT and BIER-BioBERT) perform near the level of their dense, non-interpretable counterparts. In the next section we will look at some illustrative test examples cases along with a simple technique to leverage both the dense and sparse embeddings that a BIER model can give to improve performance on the task and gain insight into where the entity type model and system may be underperforming.

Table 4 shows the results of our fine-tuning experiment. Freezing the model and allowing only the classification layer to learn weights doesn't allow enough capacity for either case, while fully fine-tuning both models gives improved performance in both models. However because the BIER model is no longer tied in, the interpretability component of our representations is eliminated, a limitation left for future work.

Figure 3 shows BIER-PubMedBERT performs better than the fine-tuned and non-interpretable PubMedBERT model when there are fewer than 100 examples per class (which is the case for 6 out of the 16 test classes in the dataset as seen in table 7 in the appendix).

	Test Acc.	
Model	Frozen Model	Fine-Tuned
BIER-PubMedBERT (ours)	68.0	96.0
PubMedBERT	36.2	96.1

Table 4: Test results on Cancer Genetics task with fine tuning on all data whether freezing the model or not.

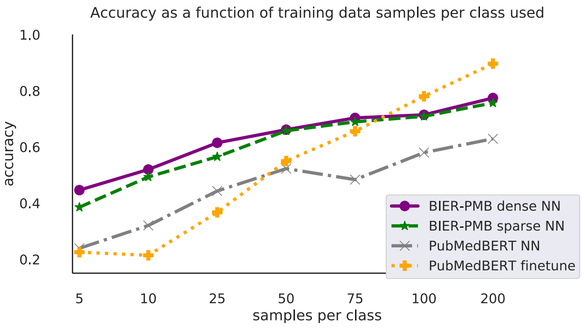


Figure 3: Results for the entity label classification task under varying amounts of supervision.

5 Debugging with BIERs

One of the claimed advantages of BIERs is their ability to facilitate model debugging. In this section we provide illustrative examples where the interpretability of the underlying representations offers insights into model behavior and suggests avenues for improvements.

Entity Type and Mention Analysis We illustrate the debugging strategy proposed in the context of entity label classification. Recall that this entails inspecting test examples for which the dense model yields a correct prediction, while the sparse variant does not (implying that the former somehow better represents the instance). We can inspect these cases to understand what entity types are leading to such behavior. Appendix Table 11 and 12 enumerate such mentions and their most probable types. We note the inclusion of many people's names (e.g., "Anthony Campbell", "Tony Walsh") which have been assigned at least some incorrect types in their sparse representations. This highlights a general failure mode of the model: It is assigning incorrect types to person names, which may be causing downstream prediction errors. This is actionable information, as we could remedy the issue via rules, additional, targeted supervision or by down weighting probabilities given to common erroneous types for these mentions.

To better characterize entity type errors, we gather the set of the 20 most probable entity types for all mentions incorrectly predicted by the BIER sparse model and sort types by frequency. We do the same for those predicted correctly. The resulting two lists share many of the same top popular types, but looking at relative rankings and only displaying those that are comparatively far apart[10] reveals some interesting results. Tables 9 and 10 report entity types correlated with correct and incorrect predictions, respectively. We emphasize this type of analysis is only possible due to the interpretable nature of the proposed BIER embeddings.

As a final illustrative debugging example, we consider a test example mention "thyroid carcinomas" with label "Cancer", along with the predictions made by the sparse model, "thyroid" with the incorrect label "Organ", and the dense model,"esophageal carcinoma" with the correct label "Cancer". We also retrieve the first correct prediction from the nearest neighbors of the sparse model embedding "medullary thyroid carcinoma" which we refer to as the *counterfactual* sparse prediction.[11] We take the dot product of the mention-context embedding with these three prediction's embeddings and inspect the top types which lead to their selection in Figure 6 in Appendix C. Both the incorrect sparse and correct counterfactual sparse predictions, at the surface level are quite similar to the test mention, but have lower scores for the entity type 'thyroid cancer' compared with the dense prediction which gives the correct label, but is semantically less similar to the test mention than the counterfactual sparse prediction. Additionally, the noisy type "rtt" erroneously plays more of a role in the sparse model predictions as well.

Diagnosing task results In analyzing errors made by the highest performing BIER dense and sparse nearest neighbor models for the entity label classification task, we noticed that while there was high concurrence for correct predictions (i.e., of the 88% true predictions made by the dense model overall, the sparse model agreed with the prediction 95% of the time), the cases where the model predictions disagreed, but where one of them still predicted the true label, were quite varied. In other words, the sparse model gave many correct results on test cases when the dense model gave incorrect ones and vice versa. Applying the diagnostic technique from Section 3, we see the classifier's overall performance could have improved from 88.2 to 91.9 had the model known when to utilize its intermediate dense representation over its sparse output.

	Test Acc.			
Task	Dense	Sparse	Combined	Δ
NED	84.0	81.0	**91.7**	+7.7
ELC	87.5	88.2	**91.9**	+3.7

Table 5: Results for both tasks showing improvements that could have been achieved by combining intermediate dense and interpretable sparse output embeddings generated by the same BIER-PubMedBERT model.

Similarly we applied the diagnostic technique to the NED task and leave more details in Appendix B. Incorporating mentions that the dense dot product BIER model handles better than the cosine similarity based sparse one does would have given an improvement from our prior accuracy of 84.0 to 91.7. Table 5 shows the possible improvement in task accuracies for both tasks.

6 Related Work

In this work we have introduced a predefined fine-grained biomedical type system comprising 68k types, explicitly tied to PubMed. Instead of using a fixed type system, Raiman and Raiman (2018) seek to dynamically learn a 100 dimensional type system from a much larger general domain type system in order to optimally disambiguate entities.

Aside from work on biomedical NLP and entities specifically, there exists a line of work on interpretable word embeddings (Subramanian et al., 2017; Faruqui et al., 2015). A common approach here is to identify the groups of words most associated with vector components globally, somewhat akin to topic models. This differs from our approach, which is based on an external type system and provides immediate, instance-level interpretable probabilities for each entity type. Hu et al. (2020) proposes transforming dense to sparse representations independent of entity typing.

Another related line of work tests a models' ability to induce syntactic or type information by the measuring accuracy of a probe (Peters et al., 2018; Hewitt and Manning, 2019; Hewitt and Liang, 2019). There is significant uncertainty about how

[10] We chose to highlight entity types that are farther than 50 rankings apart to have a small set to display.

[11] Had the mention under consideration instead mapped to this sparse representation, the prediction would have been different, and correct.

to calibrate such post-hoc probing results (Voita and Titov, 2020) whereas our model's representations are directly interpretable.

While many interesting biomedical entity representation and linking task oriented works (Murty et al., 2018; Vashishth et al., 2020; Mondal et al., 2019; Sung et al., 2020; Liu et al., 2020) leverage PubMed or UMLS for semantic type, entity synonym, or self alignment purposes, our work is the first to incorporate interpretable embeddings that are linked to a biomedical entity type system.

7 Conclusions

We have introduced a new biomedical entity typing system and training set from a large corpus of biomedical texts. We will release this dataset, which comprises 37 million derived triples. Exploiting this data, we proposed *Biomedical Interpretable Entity Representations* (BIERs), in which dimensions correspond to fine-grained entity types, and values are predicted probabilities that a given entity is of the corresponding type.

Using two downstream biomedical tasks, we showed that BIER representations yield predictive performance that is competitive with dense (uninterpretable) representations, and that such representations are particularly beneficial in zero-shot or low-supervision settings. We also demonstrated that BIER representations can facilitate meaningful model debugging both at the mention and entity type level.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation (under Grants No. NSF 1901117 and NSF 1750978) and in part by a grant from the Office of Naval Research (ONR N00014-17-1-2143). The project was conducted under the auspices of the IBM Science for Social Good initiative.

Ethical Considerations

NLP models are increasingly used in biomedicine, where some applications can be quite high-stakes. Establishing trust in such models is therefore paramount; unfortunately, deep neural networks tend to be opaque in their operations, potentially precluding their use in certain areas of biomedicine where they might otherwise be beneficial. This work is a step towards more transparent models for biomedical NLP.

References

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A Corpus for Medical Entity Linking in the Social Media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mingda Chen, Zewei Chu, Yang Chen, Karl Stratos, and Kevin Gimpel. 2019. EntEval: A holistic evaluation benchmark for entity representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-Fine Entity Typing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

John Cuzzola, E. Bagheri, and Jelena Jovanovic. 2018. UMLS to DBPedia Link Discovery Through Circular Resolution. *Journal of the American Medical Informatics Association*, 25:819–826.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. 2015. Sparse Overcomplete Word Vector Representations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing.

John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Wenpeng Hu, Mengyu Wang, Bing Liu, Feng Ji, Jinwen Ma, and Dongyan Zhao. 2020. Transformation of Dense and Sparse Text Representations. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734*.

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*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment Pre-training for Biomedical Entity Representations.

Sunil Mohan and Donghui Li. 2019. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. *CoRR*, abs/1902.09476.

Ishani Mondal, Sukannya Purkayastha, Sudeshna Sarkar, Pawan Goyal, Jitesh Pillai, Amitava Bhattacharyya, and Mahanandeeshwar Gattu. 2019. Medical Entity Linking using Triplet Network. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*.

Shikhar Murty, Patrick Verga, Luke Vilnis, Irena Radovanovic, and Andrew McCallum. 2018. Hierarchical Losses and New Resources for Fine-grained Entity Typing and Linking. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*.

Yasumasa Onoe and Greg Durrett. 2019. Learning to Denoise Distantly-Labeled Data for Entity Typing. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Yasumasa Onoe and Greg Durrett. 2020. Interpretable Entity Representations through Large-Scale Typing. In *Findings of the Association for Computational Linguistics: EMNLP*.

Nicolas Papernot and P. McDaniel. 2018. Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. *ArXiv*, abs/1803.04765.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. BERT is Not a Knowledge Base (Yet): Factual Knowledge vs. Name-Based Reasoning in Unsupervised QA. *ArXiv*, abs/1911.03681.

Sampo Pyysalo, Tomoko Ohta, and Sophia Ananiadou. 2013. Overview of the Cancer Genetics (CG) task of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*.

Jonathan Raiman and Olivier Raiman. 2018. DeepType: Multilingual Entity Linking by Neural Type System Evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Michael Ringgaard, Rahul Gupta, and Fernando C. N. Pereira. 2017. SLING: A framework for frame semantic parsing. *ArXiv*, abs/1710.07032.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? *ArXiv*, abs/2002.08910.

Sara Rosenthal, Ken Barker, and Zhicheng Liang. 2019. Leveraging Medical Literature for Section Prediction in Electronic Health Records. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Valentin I. Spitkovsky and Angel X. Chang. 2012. A Cross-Lingual Dictionary for English Wikipedia Concepts. In *Proceedings of LREC*.

Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard H. Hovy. 2017. SPINE: Sparse Interpretable Neural Embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical Entity Representations with Synonym Marginalization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

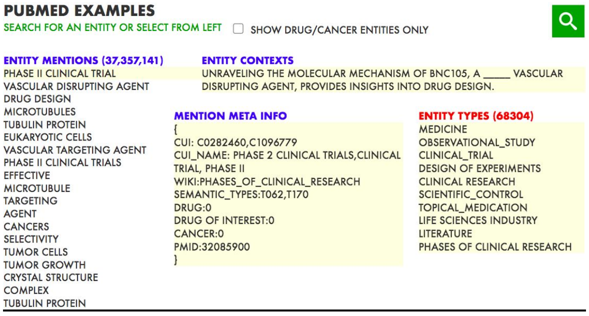
Shikhar Vashishth, Rishabh Joshi, Denis Newman-Griffis, Ritam Dutt, and Carolyn Rose. 2020. MedType: Improving Medical Entity Linking with Semantic Type Prediction.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Elena Voita and Ivan Titov. 2020. Information-Theoretic Probing with Minimum Description Length. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv*, abs/1609.08144.



PUBMED EXAMPLES

SEARCH FOR AN ENTITY OR SELECT FROM LEFT □ SHOW DRUG/CANCER ENTITIES ONLY

ENTITY MENTIONS (37,357,141)

PHASE II CLINICAL TRIAL
VASCULAR DISRUPTING AGENT
DRUG DESIGN
MICROTUBULES
TUBULIN PROTEIN
EUKARYOTIC CELLS
VASCULAR TARGETING AGENT
PHASE II CLINICAL TRIALS
EFFECTIVE
MICROTUBULE
TARGETING
AGENT
CANCERS
SELECTIVITY
TUMOR CELLS
TUMOR GROWTH
CRYSTAL STRUCTURE
COMPLEX
TUBULIN PROTEIN

ENTITY CONTEXTS

UNRAVELING THE MOLECULAR MECHANISM OF BNC105, A _____ VASCULAR DISRUPTING AGENT, PROVIDES INSIGHTS INTO DRUG DESIGN.

MENTION META INFO

```
{
CUI: C0282460,C1096779
CUI_NAME: PHASE 2 CLINICAL TRIALS,CLINICAL
TRIAL, PHASE II
WIKI:PHASES_OF_CLINICAL_RESEARCH
SEMANTIC_TYPES:T062,T170
DRUG:0
DRUG OF INTEREST:0
CANCER:0
PMID:32085900
}
```

ENTITY TYPES (68304)

MEDICINE
OBSERVATIONAL_STUDY
CLINICAL_TRIAL
DESIGN OF EXPERIMENTS
CLINICAL RESEARCH
SCIENTIFIC_CONTROL
TOPICAL_MEDICATION
LIFE SCIENCES INDUSTRY
LITERATURE
PHASES OF CLINICAL RESEARCH

Figure 4: Example from derived Biomedical Dataset

A BIER system level specifics

To better illustrate the process of which mentions are retained during our filtering process, Table 6 shows the 6 concepts associated with an example mention of ”phase II clinical trial” found in a PubMed article. We see all 6 concepts score higher than our minimum threshold and we use the two highest scoring matches that are within 2 points of each other: CUIDs C0282460 and C1096779. the former C0282460 has a WikiPedia data item Q7180990 that corresponds to the page “wiki/Phases_of_clinical_research” whose associated categories are “Clinical research”, “Design of experiments”, “Life sciences”, “industry”. The second result C1096779 has no direct WikiPedia match and the results we get from SLING include “Clinical trial”, “Scientific control”, “Medicine”, “Topical medication”, “Observational study”, “Literature”. Hence for this mention and context from a PubMed abstract, we are able to extract a (mention, context, list of types) triple of the form (“phase II clinical trial”, context, [“Clinical research”, “Design of experiments”, “Life sciences” industry”, “Clinical trial”, “Scientific control”, “Medicine” ....]].

CUID	Concept Name	Score	DBPedia
C0282460	Phase 2 Clinical Trials	0.9999	Q7180990
C1096779	Clinical Trial, Phase II	0.9999	none
C0282461	Phase 3 Clinical Trials	0.9496	Q7180990
C0920321	Phase I Clinical Trials	0.8707	Q7180990
C1096780	Clinical Trial, Phase III	0.8635	none
C0282462	Phase 4 Clinical Trials	0.8208	Q7180990

Table 6: Using an NER tagger we find 6 associated concepts in UMLS for the mention “phase II clinical trial” in a context sentence “Unraveling the molecular mechanism of BNC105, a phase II clinical trial vascular disrupting agent, provides insights into drug design.”

Label	Train + Dev Set % (raw)	Test Set % (raw)
Gene_or_gene_product	36.98 (5388)	36.23 (2520)
Cell	17.32 (2524)	15.15 (1054)
Cancer	11.52 (1679)	13.30 (925)
Simple_chemical	10.59 (1543)	10.45 (727)
Organism	8.63 (1258)	7.81 (543)
Multi-tissue_structure	3.80 (554)	4.36 (303)
Tissue	2.77 (403)	2.73 (190)
Cellular_component	2.67 (389)	2.59 (180)
Organ	1.82 (265)	2.24 (156)
Organism_substance	1.24 (181)	1.47 (102)
Pathological_formation	0.96 (140)	1.28 (89)
Amino_acid	0.50 (73)	0.89 (62)
Immaterial_anatomical_entity	0.49 (71)	0.45 (31)
Organism_subdivision	0.40 (59)	0.56 (39)
Anatomical_system	0.16 (24)	0.24 (17)
Developing_anatomical_structure	0.12 (18)	0.24 (17)

Table 7: Cancer Genetics Dataset Label Distribution

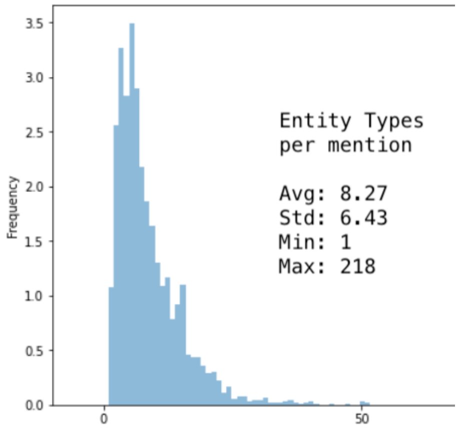


Figure 5: Entity Types per mention on Training set for BIER

B NED diagnostic details

For the NED task we used the BIER’s sparse embeddings of test mentions in their contexts and took cosine similarity with a separate BIER model’s sparse embeddings of candidate wiki descriptions to make our predictions. To use the diagnostic technique we first get task predictions using the dense embeddings from the BIER models which gives results of 81 and 79.25 percent test accuracy using dot product and cosine similarity respectively. Although the prior sparse cosine similarity BIER model in this case gave a higher 84.0 percent test accuracy, using the diagnostic technique in this case by incorporating mentions the dense dot product BIER model handles better would have given an improvement in accuracy from 84.0 to 91.65.

1-25	26-50	51-75	76-100
term	w.h.o. essential medicines	test_(assessment)	causality
ingredient	psychosis	drug discovery	hydroxyl
disease	scientific_method	receptor_(biochemistry)	adverse_effect
cell_(biology)	oncology	observational_study	diagnosis
rtt	enzyme	immunology	physiology
protein	human_body	molecular biology	chemotherapy
gene	psychotherapy	abnormality_(behavior)	hepatotoxins
human	medicine	radioactive_decay	molecule
neoplasm	grammatical_modifier	derivative_(chemistry)	phenotype
cancer	tissue_(biology)	chemistry	cell biology
therapy	treatment_and_control_groups	health policy	concepts in metaphysics
medical terminology	scientific method	amine	concepts in epistemology
measurement	coagulation	peptide	apoptosis
patient	chemical_reaction	pharmaceutical sciences	procedural_law
chemical_compound	philosophy of science	antigen	science
surgery	calcium_in_biology	biology	genetic_code
nitrous_oxide	enzyme_inhibitor	algorithm	empiricism
pharmaceutical_drug	medicinal chemistry	texas	family
acid	research	mental_disorder	thailand
articles containing video clips	metabolism	statistical_hypothesis_testing	liver
malignancy	taxonomy_(biology)	catalysis	medical mnemonics
time	cell_growth	allele	dosage_form
prothrombin_time	blood	methyl_group	immune system
cognition	syndrome	infectious causes of cancer	amino_acid
drug	sewage_treatment	database	beta_sheet

Table 8: Top 100 most frequent types from Biomedical Entity Type System

Incorrect rank	Entity Type	Correct rank	Relative difference
20	tongue	76	56
24	anatomy	160	136
29	protein_domain	112	83
34	organ_(anatomy)	107	73
35	gland	205	170
38	phosphatase	140	102
43	surgery	120	77
46	circulatory_system	293	247
50	squamous-cell_carcin	111	61
51	nephron	142	91
60	anatomical_terms	169	109
61	kidney	284	223
62	cancer_cell	213	151
70	activator_(genetics)	179	109
71	drug	192	121
74	breast_cancer	127	53
75	locus_(genetics)	206	131
77	cancer_staging	256	179
79	signal transduction	233	154
81	multiprotein_complex	132	51
82	endometrium	200	118
83	mouth	200	117
84	cell anatomy	272	188
90	molecular biology	200	110
93	rare cancers	200	107
95	website	161	66
96	cell_cycle	200	104
97	gene expression	178	81
98	hydroxyl	221	123
99	oral_sex	200	101

Table 9: Entity Types more associated with erroneous predictions

Incorrect rank	Entity Type	Correct rank	Rel diff
23	syndrome	244	221
34	abnormality_(behavior)	270	236
35	elementary_particle	128	93
42	apoptosis	200	158
51	congenital_disorder	200	149
53	transformation_(genetics)	275	222
54	measurement	109	55
55	human cells	147	92
56	immune system disorders	200	144
57	paraneoplastic syndromes	200	143
58	code	154	96
59	battery_(electricity)	200	141
61	virus	222	161
63	chemistry	281	218
66	calcium_in_biology	209	143
71	thymus	200	129
72	medical terminology	190	118
73	cell biology	297	224
74	recombinant_dna	10	64
76	tongue	20	56
79	protein_kinase	164	85
80	drama	200	120
85	tumor_suppressor_gene	14	71
86	patient	200	114
87	specialty_(medicine)	234	147
90	growth_hormone	200	110
91	taxonomy_(biology)	238	147
93	t cells	228	135
94	childhood	200	106
95	aging-related proteins	200	105
96	network_affiliate	200	104
97	blood tests	200	103
98	protein_a	200	102

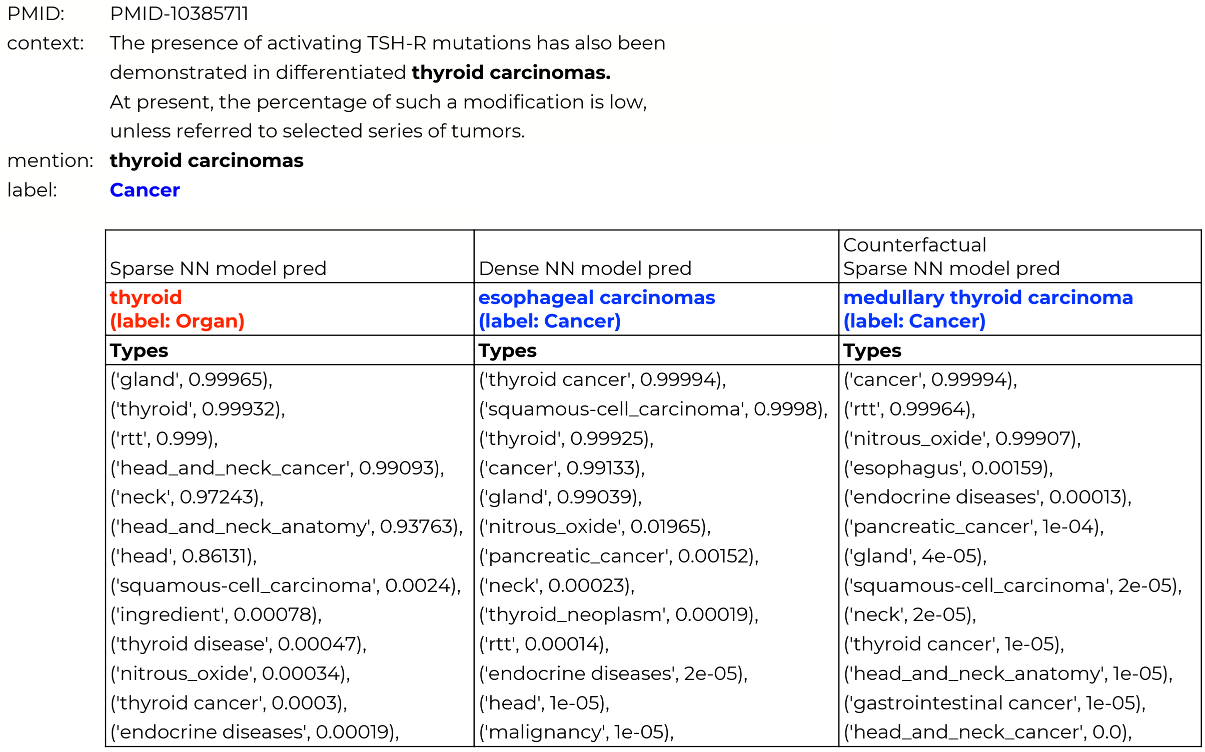
Table 10: Entity Types more associated with correct predictions

Mention	Sparse embedding top types that do worse than dense counterparts
albinism	albinism, disease, animal coat colors, heredity, dermatologic terminology, articles containing video clips, hair, skin, pigment, nitrous_oxide
alveolar ridge	tooth, lung anatomy, mouth, pulmonary_alveolus, human mouth anatomy, dental_caries, periodontology, parts of tooth, mandible, leaf
anastomosis	surgery, anastomosis, evolutionary biology, digestive system, angiology, anatomy, lawsuit, combat, organ_(anatomy), surgical_anastomosis
anthony campbell	carl_linnaeus, ingredient, human, taxa named by carl linnaeus, flora of asia, world health organization essential medicines, coordination_complex, western european countries, extract, asteraceae genera
autonomic neuropathy	peripheral nervous system disorders, autonomic nervous system, disease, nervous_system, peripheral_neuropathy, functional_group, heredity, mental_disorder, nerve
bleb	skin conditions resulting from physical factors, lesion, fluid, frostbite, radiation health effects, disease, source_code, nitrous_oxide, skin, hematology
cancer cells	cell_(biology), cell_culture, medical terminology, oncology, cancer, precancerous_condition, large_cell, human, protein, standard_operating_procedure
cell death	mitochondria, programmed cell death, death, cell_(biology), cellular senescence, cognition, apoptosis, tgf_beta_signaling_pathway, nuclear_receptor, survival_rate
chronic lymphocytic leukemia	plasma_cell, small_intestine, mood_disorder, b-cell_lymphoma, lymphocytic leukemia, bioaccumulation, bone_marrow, grading_(tumors), lymphatic_system, lymphoblast
cosmological constant	ionizing radiation, assumption, units_of_measurement, comics by steve ditko, cell_(biology), grammatical_modifier, quantity, blood_plasma, industrial gases, litre
demethylation	organic reactions, gene expression, posttranslational modification, rna, transcription_(genetics), molecular genetics, epigenetics, therapy, demethylation, molecular biology
dissociation constant	wine regions of south africa, suburbs of cape town, astronomical_unit elementary_particle, rat, medical terminology, gene, units_of_measurement, furans
endoscope	endoscopy, bicycle, diagnostic gastroenterology, physical_examination, gastroenterology, microphone, video_camera, israeli inventions, pencil, 21st-century inventions
fingering	finger, conditions of the skin appendages, articles containing video clips, disease, toe, nitrous_oxide, hand, keratin, reflex, fingers
flirting	female, causes of death, fly, metrorrhagia, disease, articles containing video clips, conditions of the skin appendages, etiology, dog, homology_(biology)
geniculate	tongue, anatomical_terms_of_location, ganglion, mandible, midbrain, organ_(anatomy), cell_nucleus, cerebral_cortex, lobe_(anatomy), middle_ear
glomerulus	kidney, tongue, connective_tissue, epithelium, gene, cell_membrane, nitrous_oxide, organ_(anatomy), nephrology, derivative_(chemistry)
guy davis	hemoglobins, respiratory physiology, hemoglobin, geography, equilibrium chemistry, cancer, race_and_ethnicity_in_the_united_states_census, geographic_coordinate_system, texas
infant formula	infant, infant feeding, child, milk, formula, dosage_form, foods, breast milk, preterm_birth, chemistry
intussusception	bowel_obstruction, human_gastrointestinal_tract, large_intestine, invagination, disease, morphology_(biology), colorectal_cancer, nitrous_oxide, deconstruction, thrombosis
isomerization	isomerism, stereochemistry, metabolism, 1827 introductions, laboratory techniques, transgender, chemistry, organic chemistry, isomerases, flora of california
mescaline	ingredient, rtt, mesylate, abbvie inc. brands, anti-inflammatory, orphan drugs, acid, methyl_group, carbamates, amine
methylation	epigenetics, posttranslational modification, amino_acid, protein, methylation, acid, organic reactions, amine, antigen, ingredient

Table 11: NED examples where dense BIER embeddings outperforms sparse (interpretable) BIER representations. Mentions start with [A-M].

Mention	Sparse embedding top types that do worse than dense counterparts
n400	antigen, cancer, gene, protein, units_of_measurement, ratio, allele, human, time, nucleolus
peroxides	acne_vulgaris, topical_medication, ingredient, functional_group, route_of_administration, peroxides, chemical_reaction, glandular and epithelial neoplasia, functional groups, pharmaceutical_drug
polyunsaturated fatty acids	nutrition, fatty acids, acid, lipids, ester, protein, ingredient, food science, neuronal_ceroid_lipofuscinosis, lipid
protease inhibitors	endopeptidase, enzyme_inhibitor, biosynthesis, protease, chemical_compound, receptor_antagonist, peptide, moa, hiv, hiv-1_protease
psychological dependence	psychology, substance_dependence, substance_abuse, emotion, mental_disorder, dependent territories, governance of the british empire, mental and behavioural disorders, crown dependencies, british islands
psychopathy	abnormal psychology, abnormality_(behavior), nitrous_oxide, pathology, disease, behavioural sciences, psychosis, mental and behavioural disorders, mental_disorder, affect_(psychology)
resection	technology, natural_resource, segmental_resection, plant anatomy, plant physiology, surgical_suture, surgery, plant morphology, morphology_(biology), amputation
sarcoidosis	epithelioid_cell, etiology, chilblains, disease, nitrous_oxide, kalashnikov derivatives, sarcoidosis, organ_(anatomy), 5.56×45mm nato assault rifles, carbines
semicircular canals	tongue, ear_canal, vestibular system, auditory system, eustachian_tube, canal_(anatomy), auditory_system, crystal_structure, vestibulocochlear nerve, cranial_cavity
sequence analysis	sequence, bioinformatics, psychosis, psychoanalysis, dna, scientific method, nucleic_acid_sequence, physical_examination, dna_sequencing, algorithm
tony walsh	race_and_ethnicity_in_the_united_states_census, adult, flora of asia, carl_linnaeus, human, french-speaking countries, flora of north america, hemoglobins, women, coagulation system
united states department of agriculture	united_states, united states federal executive departments, management, police, public_health, united_states_department_of_defense, united states department of health and human services agencies, regulators of biotechnology products, 1889 establishments in the united states
ventricular zone	ventricle_(heart), zoning, ventricular_system, ventricular system, brain, developmental neuroscience, tongue, urban planning, anatomical_terms_of_location, bone
wechsler adult intelligence scale	psychological testing, psychiatric assessment, connective/soft tissue tumors and sarcomas, nitrous_oxide, psychiatric diagnosis, medical scales, level_of_measurement, adult, childhood
yang xiong	yin_and_yang, qi, alternative medicine, taoist cosmology, chinese martial arts terminology, chinese philosophy, plants used in traditional chinese medicine, gene, qigong, trees of china

Table 12: NED examples where dense BIER embeddings outperform sparse BIER representations. Mentions start with [N-Z].



PMID: PMID-10385711
context: The presence of activating TSH-R mutations has also been demonstrated in differentiated **thyroid carcinomas.** At present, the percentage of such a modification is low, unless referred to selected series of tumors.
mention: **thyroid carcinomas**
label: **Cancer**

Sparse NN model pred	Dense NN model pred	Counterfactual Sparse NN model pred
thyroid **(label: Organ)**	**esophageal carcinomas** **(label: Cancer)**	**medullary thyroid carcinoma** **(label: Cancer)**
Types	**Types**	**Types**
('gland', 0.99965), ('thyroid', 0.99932), ('rtt', 0.999), ('head_and_neck_cancer', 0.99093), ('neck', 0.97243), ('head_and_neck_anatomy', 0.93763), ('head', 0.86131), ('squamous-cell_carcinoma', 0.0024), ('ingredient', 0.00078), ('thyroid disease', 0.00047), ('nitrous_oxide', 0.00034), ('thyroid cancer', 0.0003), ('endocrine diseases', 0.00019),	('thyroid cancer', 0.99994), ('squamous-cell_carcinoma', 0.9998), ('thyroid', 0.99925), ('cancer', 0.99133), ('gland', 0.99039), ('nitrous_oxide', 0.01965), ('pancreatic_cancer', 0.00152), ('neck', 0.00023), ('thyroid_neoplasm', 0.00019), ('rtt', 0.00014), ('endocrine diseases', 2e-05), ('head', 1e-05), ('malignancy', 1e-05),	('cancer', 0.99994), ('rtt', 0.99964), ('nitrous_oxide', 0.99907), ('esophagus', 0.00159), ('endocrine diseases', 0.00013), ('pancreatic_cancer', 1e-04), ('gland', 4e-05), ('squamous-cell_carcinoma', 2e-05), ('neck', 2e-05), ('thyroid cancer', 1e-05), ('head_and_neck_anatomy', 1e-05), ('gastrointestinal cancer', 1e-05), ('head_and_neck_cancer', 0.0),

Figure 6: Analysis Example for ELC task