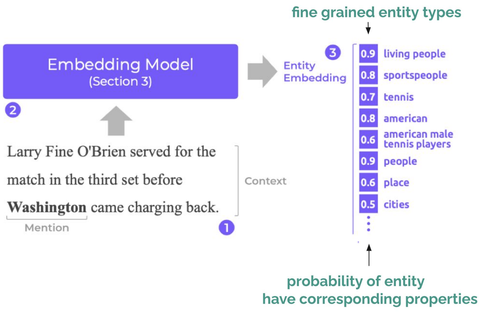
ItsIRL: Intermediate Entity-based Sparse Interpretable Representation Learning

Diego Garcia-Olano, Yasumasa Onoe, Joydeep Ghosh, Byron Wallace

BlackboxNLP 2022 at EMNLP
Workshop on analyzing and interpreting neural networks for NLP

Background and Motivation

Interpretable entity representations (IERs) are pre-trained sparse "human-readable" embeddings whose dimensions correspond to fine-grained entity types and values are the predicted probabilities that a given entity is of a corresponding type from a large, static type system.



IERs are inherently interpretable without the need for probing and are used directly for classification.

They **perform well in zero-shot and low supervision** settings (*Onoe, 2020; Garcia-Olano 2021*) and compared with standard dense embeddings, IERs permit **component level analysis and debugging.**

However, while fine-tuning IERs improves accuracy on downstream tasks, it destroys the semantics of the dimensions which were enforced in pre-training.

Question: Can we maintain the interpretable semantics afforded by IERs while improving predictive performance on downstream tasks?

Our proposal: ItsIRL

We propose Intermediate enTity-based Sparse Interpretable Representation Learning (ItsIRL). See Fig 1.

• We **introduce an intermediate interpretable layer** into IERs; this layer output is then "decoded" into a dense layer which can be used for downstream predictions. This **decoding step can be fine-tuned** for specific tasks.

• We show that this approach **empirically outperforms prior IER methods** on two diverse biomedical benchmark tasks, often by a substantial margin.

• We propose a **counterfactual entity type manipulation analysis** made possible by our architecture which facilitates **model debugging** in an automated fashion with minimal, noisy supervision. This analysis allows our model to outperform dense (uninterpretable) models in terms of test accuracy and shows that the entity typing layer affects output classifications in an intuitive way.

• We create **positive and negative class prototypes combining entity types** over classes on the training set to **reveal global semantics** learned by our model.

Experimental Results

Task: Cancer Genetics Entity Label Classification
& Ablation Studies for Initialization & Decoder Size

Model	🔍	Test Acc
BIER-PMB*	✓	87.5
ItsIRL	✓	91.9
ItsIRL E2E*	-	95.7
PubMedBERT	-	96.1

Ablations	Test Acc
ItsIRL - random init	88.9
ItsIRL - 1 layer decoder	68.1

Table 1: Cancer Genetics results
🔍 = interpretable types

Code: https://github.com/diegoolano/itsirl

ItsIRL Architecture

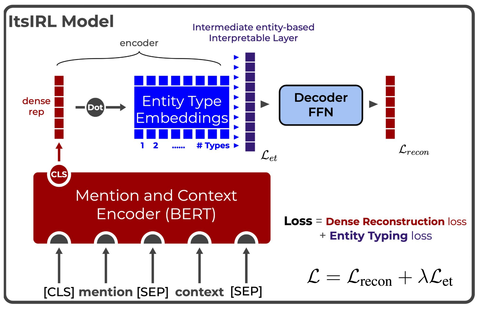


Figure 1: ItsIRL uses a LM and type supervision during pre-training to encode entity mention and context inputs for learning a matrix of entity type embeddings, an intermediate interpretable layer of type scores and a decoder to reconstruct the initial LM representation. The decoder can be fine-tuned on downstream tasks for better performance than IERs while keeping the semantics of the type layer.

Motivation behind the additional reconstruction loss is to pre-train a sort of auto-encoder with a sparse, high dimensional, interpretable latent space and rich dense output representations. The encoder induces a sparse embedding of entity types as in prior work on IERs, but now for downstream tasks we can freeze the encoder (which yields interpretable entity representations) and *fine-tune* the decoder. This allows for both interpretable entity types and improved task performance.

Counterfactual Type Manipulation

We explore how intermediate entity types affect task performance and ***how predictions would have changed*** had relevant types been manipulated.

- We construct sets of entity types for each class
- We propose 3 strategies for manipulating entity types during inference time and observe how final class probabilities for the task are affected by each.

Model	Test Accuracy
ItsIRL	91.48
+ Fix types	93.91
+ Promote types	95.74
+ Both fix & promote	95.68
+ Best of 3 "oracle"	**96.78**
PubMedBERT*	96.10

Entity type manipulation results using class-specific coarse type sets

Entity Type Class Based Prototypes

To better understand the representations learned by ItsIRL for each class, we apply the task, decoder fine-tuned model over the training data.

We gather all correctly predicted instances for each class, sum their interpretable entity type representations and normalize them. These are ***positive class prototypes.***

	Gene or gene product	Cell	Cancer	Simple chemical
1	protein	cell	disease	ingredient
2	ingredient	elementary particle	neoplasm	acid
3	human	human cells	oncology	rtt
4	gene	battery	tissue	who essential medicines
5	coagulation	gene	abnormality	chemical compound
6	cell	protein	cancer	measurement
7	cell growth	pancreas	syndrome	calcium

Top 7 Entity Types by weight for 4 most frequent positive Prototype class embeddings

Task: BIOSSES sentence similarity regression
& Entity Type Sparsity Analysis

			Type Sparsity		
Model	🔍	MSE	@.01	@.1	@.25
BIER-PMB*	✓	5.05	33.6	8.1	4.4
ItsIRL	✓	1.59	33.6	8.1	4.4
ItsIRL E2E*	-	1.15	5723	780	330
PubMedBERT	-	1.14	-	-	-

Table 2: BIOSSES sentence similarity results.

PMB* = PubMedBERT
E2E* = End-To-End fine-tuned

Take a photo to learn more:

References

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

Diego Garcia-Olano, Yasumasa Onoe, Ioana Baldini, Joydeep Ghosh, Byron Wallace, and Kush Varshney. 2021. Biomedical interpretable entity representations. In Findings of the Association for Computational Linguistics: EMNLP.

TEXAS The University of Texas at Austin
Northeastern University
Meta AI