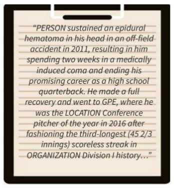
Mixture of Hierarchical Unified Neural Domain Experts (MHUNDE): Transforming Scouting in Major League Baseball

Amol Singh[1], Aman Malhotra[1], Eish Maheshwari[1]

[1]Department of Computer Science, Stanford University

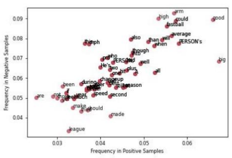
Problem

- The majority of baseball prospects do not make it to the major leagues
- Essential task: to effectively identify talent in the prospect pool
- We aim to implement an effective pre-trained and fine-tuned deep learning model to predict whether a baseball prospect will have a major league career, given scouting reports written on the player.



Background

A labeled scouting report dataset and simple baselines on the binary classification task were provided by Jacob Danovitch and are shown in Figure 1b:



Model	Accuracy	F1
Bag-Of-Embeddings	64.65%	53.78%
TextCNN	69.02%	56.42%
LSTM+SelfAttn	68.64%	54.65%
BCN	73.52%	43.33%
HAN	66.00%	54.07%

Fig 1: a) Frequency scatter plot excluding top 40 overlaps which consisted of mostly prepositions and common nouns. This scatter plot indicates the difficulty of this problem as both negative and positive scouting reports use very similar vocabulary but within different grammatical contexts. **b)** The results reported by Donovitch 2019 as a result of training a series of different models to predict the MLB prospects based on the labeled scouting reports.

The main issues we found with existing modeling approaches are: small dataset, class imbalance, niche domain.

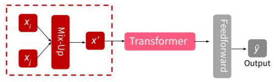
The goal of our project is to address these issues with a three-pronged approach:

1) Data augmentation
2) Domain-adaptive pre-training (DAPT) and task-adaptive pre-training (TAPT)
3) Mixture of Experts (MOE)

We propose a mixture of hierarchical unified neural domain experts (MHUNDE) as think tanks, where each "expert" is trained with domain-adaptive pretraining (DAPT) or task-adaptive pretraining (TAPT) on a BERT base.

Methods

Data Augmentation



$$x' = \lambda x_i + (1-\lambda)x_j$$
$$y' = \lambda y_i + (1-\lambda)y_j$$

Fig 2: Architecture of the hybrid mix-up transformer. x_i and x_j are tokenized embeddings of 2 scouting reports. Inputs are linearly interpolated before fed through a pre-trained transformer (ex: BERT) and feedforward layer with dropout. BERT weights are unfrozen during fine-tuning.

DAPT/TAPT

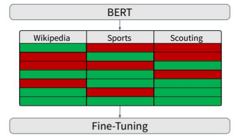


Fig 3: Architecture of hierarchical pre-training framework. A BERT transformer is pre-trained on a combination of Wikipedia articles, general sports articles, and unlabeled scouting reports (using a masked language modeling objective) before fine-tuned on labeled scouting reports for binary classification.

Mixture of Domain Experts

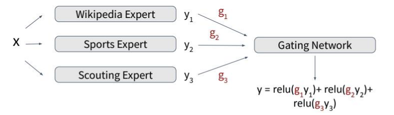


Fig 4: Architecture of the MOHE framework. Predictions from individual domain experts (each well-versed in their respective domains through targeted pre-training) are passed through a learned gating network (3-layer fully connected NN) to produce a final ensembled prediction.

Experiments

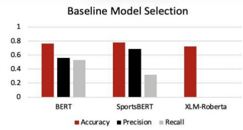


Fig 5: Performance of various baseline models without pretraining or fine-tuning (best: BERT)

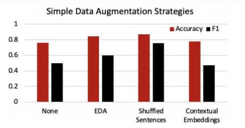


Fig 6: Performance of various data augmentation strategies models (best: shuffling sentences)

MixUp Transformer

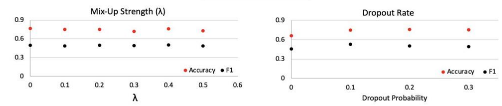


Fig 7: Hyperparameter tuning of mix-up transformer. Ideal performance was with mixup strength of $\lambda = 0.4$ and dropout rate of $p = 0.2$ with accuracy = 75.6% and F1 = 0.5007. Overall, there was little improvement with the mixup transformer.

Hierarchical Pre-Training

Pre-Training	Accuracy	F1
None	*0.8689*	*0.7548*
Wiki	0.9107	0.8311
Sports	0.9067	0.8345
Scouting	**0.9357**	**0.8747**
Wiki + Sports	0.9229	0.8485
Wiki + Scouting	0.9081	0.8065
Sports + Scouting	0.9068	0.8153
Wiki+Sports+ Scout	0.9331	0.8706

Fig 8: Hierarchical pretraining on Wikipedia, sports articles, and unlabeled scouting data showed immense improvements in combination with shuffled sentences data augmentation.

Mixture of Experts

Accuracy	F1
0.9449	**0.8725**

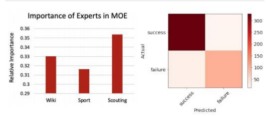


Fig 9: Mixture of domain experts showed improvement from TAPT pretrained models, with the scouting expert weighted the most (35%).

Results and Discussion

Best Model: Shuffle Sentences + Mixture of Domain Experts

- Shuffling sentences and using mixture of domain experts led to 36.9% improvement in accuracy and 54.6% increase in F1 score from the previous best model by Danovitch 2019 (textCNN)
- Biggest performance boost from shuffling sentences data augmentation (51% improvement in F1) ⟶ suggests sentence order does not matter
- Pretraining on unlabeled scouting data (TAPT > DAPT) ⟶ scouting reports & contain highly specialized jargon and patterns not captured in general articles

Mix-Up is an ineffective strategy for scouting report data

- 0.58% reduction in accuracy and only 4.14% improvement in F1 from unaugmented dataset
- Minimal effect on performance ⟶ scouting reports seem to have little in common with each other
- Tends to be recall-biased (recall > 0.9 for all models)

Mixture Of Hierarchical Experts Catches More Successes

- Recall increase by 3.4% ⟶ optimization for finding all potential successful players, ensuring that we aren't missing out on any successes

Future Work and Reference

- To reduce overfitting, develop method to introduce sentence-level order-based noise as brownian motion (BRWN-MHUNDE)
- Improve the gating function beyond a fully-connected network
- Collect more data and test transferability to different sports domains

Jacob Danovitch. Trouble with the curve: Predicting future mlb players using scouting reports, 2019

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Association for Computational Linguistics (ACL), 2020

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip Yu, and Lifang He. Mixup-transformer: Dynamic data augmentation for nlp tasks. In Salesforce Research, 2020