Exploration and Analysis of Debiasing methods for Entity Detection in Biomedical text

ADVISOR:

DR. ASHISH ANAND

By:
Abhishek Pratap Singh
MTech CSE
214101002

Problem Statement

- Generalization ability of NER models is overestimated
- The test set consists of high percentage of previously seen entities
- Overall high performance gives a false sense of generalization

Datasets	Memorized	Synonyms	Novel entities
BC5CDR	63%	28%	9%
MedMentions	70%	24%	6%
NCBI-disease	63%	20%	17%

Evaluation Framework

We are using evaluation framework from Kim et al [1].

The type of mentions is divided into three types:

1. Memorization generalization **(MEM)** - Mentions which are seen both in training as well as test set.
2. Synonym generalization **(SYN)** - Mentions whose different surface form was seen in train set
3. Concept generalization **(CON)** - These are novel entities not seen in train set

Evaluating performance across these 3 categories gives us a better idea about generalization

[1] Hyunjae Kim and Jaewoo Kang. **How do your biomedical named entity recognition models generalize to novel entities?** IEEE Access, 10:31513–31523, 2022

Datasets

The datasets used for these experiments are:

- MedMentions
- BC5CDR
- NCBI-disease

All three are available in same format.

```
25763772|t|DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cys
25763772|a|Pseudomonas aeruginosa (Pa) infection in cystic fibrosis (CF) patients
 Pa, and 68 patients of them had CPA. DCTN4 variants were identified in 24% (29/1
25763772	0	5	DCTN4	T116,T123	C4308010
25763772	23	63	chronic Pseudomonas aeruginosa infection	T
25763772	67	82	cystic fibrosis	T047	C0010674
25763772	83	120	Pseudomonas aeruginosa (Pa) infection	T047	C
25763772	124	139	cystic fibrosis	T047	C0010674
25763772	141	143	CF	T047	C0010674
25763772	145	153	patients	T101	C0030705
25763772	179	188	long-term	T079	C0443252
25763772	189	206	pulmonary disease	T047	C0024115
25763772	211	227	shorter survival	T169	C0220921
25763772	233	253	chronic Pa infection	T047	C0854135
25763772	255	258	CPA	T047	C0854135
25763772	279	300	reduced lung function	T033	C0847557
25763772	302	329	faster rate of lung decline	T033	C3160731
25763772	341	346	rates	T081	C1521828
25763772	350	363	exacerbations	T033	C4086268
25763772	368	384	shorter survival	T169	C0220921
```

Preprocessing

```
The	O
hypotensive	B
effect	O
of	O
100	O
mg	O
/	O
kg	O
alpha	B
-	I
methyldopa	I
was	O
also	O
partially	O
reversed	O
by	O
naloxone	B
.	O
```

Debiasing methods

- Reweighting [2]:
  - This method assigns weight to each example depending upon how "easy" an example is for the model to predict.

$$w(x_i) = 1 - b_i$$

where b_i is probability assigned by biased model to gold label

- Bias Product [2]:
  - Uses two model. One bias (b_i) and one main model (p_i)
  - Makes both model to learn non-overlapping features
  - Isolates bias inducing features to bias model

$$d_i = \text{softmax}(\log(p_i) + \log(b_i))$$

[2] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. **Don't take the easy way out: Ensemble based methods for avoiding known dataset biases.**(EMNLP-IJCNLP), 2019. ACL

- Learned-Mixin + H [2]:
  - Adds a learned function to the bias model output
  - Also adds an entropy penalty to the loss function makes sure the model does not make g(x) = 0.

$$d_i = \text{softmax}(\log(p_i) + g(x_i)\log(b_i))$$

$$R = wH(\text{softmax}(g(x_i)\log(b_i)))$$

[2] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. **Don't take the easy way out: Ensemble based methods for avoiding known dataset biases.**(EMNLP-IJCNLP), 2019. ACL

Bias Model

Prior Probability model:

- The probability of each token is calculated depending upon the tag distribution in train set
- if a token T1 has the tag distribution [B:3, I:4, O:2] in training set, the equivalent probability distribution will be [3/9, 4/9, 2/9] = [0.33, 0.44, 0.23]

BiLSTM-based model

- We use a single layer BiLSTM based model as bias model.

Use simpler model for bias model as it can model simple patterns only

Results

NCBI-disease + Prior Probability						
Model (early stopped)	Precision	Recall	F1-score	Mem	Syn	Con
Biobert	83.7	89.7	86.6	94.6	77.5	85.8
Bias product	82.6	87.2	84.9	91.3	79.1	82.1
Learned-Mixin+H	76.1	88.8	82.0	92.6	81.2	84.0
Reweight	85.3	89.6	87.4	94.6	79.6	82.7

NCBI-disease + BiLSTM debias						
Model (early stopped)	Precision	Recall	F1-score	Mem	Syn	Con
Biobert	83.7	89.7	86.6	94.6	77.5	85.8
Bias product	82.7	89.9	86.1	93.5	81.7	86.4
Learned-Mixin+H	70.6	87.9	78.3	91.3	80.6	84.0
Reweight	83.7	89.9	86.7	93.5	82.2	85.8

BC5CDR + Prior Probability						
Model (early stopped)	Precision	Recall	F1-score	Mem	Syn	Con
Biobert	82.9	89.5	86.1	93.9	81.3	84.4
Bias product	82.1	89.8	85.8	93.5	83.2	85.0
Learned-Mixin+H	81.6	88.2	84.8	92.5	80.4	82.7
Reweight	82.8	90.0	86.3	94.0	82.7	85.0

BC5CDR + BiLSTM debias						
Model (early stopped)	Precision	Recall	F1-score	Mem	Syn	Con
Biobert	82.9	89.5	86.1	93.9	81.3	84.4
Bias product	82.7	90.3	86.4	95.0	81.8	84.1
Learned-Mixin+H	79.1	88.7	83.6	92.7	80.9	84.4
Reweight	80.6	90.2	85.1	94.1	83.1	85.2

MedMentions + Prior Probability						
Model (early stopped)	Precision	Recall	F1-score	Mem	Syn	Con
Biobert	65.6	68.9	67.2	70.4	65.4	65.6
Bias product	62.7	68.1	65.3	69.7	64.3	64.6
Learned-Mixin+H	54.4	60.8	57.4	61.2	59.8	59.5
Reweight	63.3	68.0	65.6	70.0	63.2	63.6

MedMentions + BiLSTM debias						
Model (early stopped)	Precision	Recall	F1-score	Mem	Syn	Con
Biobert	65.6	68.9	67.2	70.4	65.4	65.6
Bias product	62.9	67.2	65.0	68.7	63.6	63.1
Learned-Mixin+H	56.2	61.5	58.7	62.9	58.2	58.6
Reweight	61.2	70.6	65.6	73.5	63.5	63.0

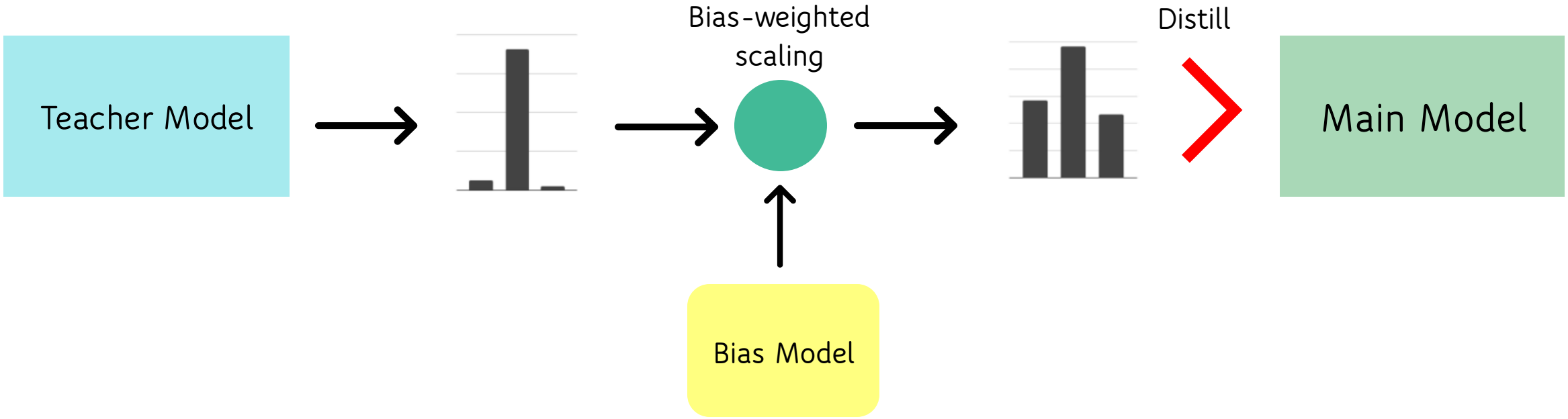
Why previous methods don't perform well for MedMentions?

The medmentions dataset has less number of biased samples.

Datasets	Percentage of tokens biased towards single tag
BC5CDR	54%
NCBI-disease	49%
MedMentions	9%

Confidence regularization [3]

- Distill a main model from teacher model
- Main model is used for inference
- The output from teacher model is scaled using a bias model



[3] Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. **Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance.** July 2020, ACL.

Scaling function of Confidence regularization

$$S(\hat{p}_i, \beta_i)_j = \frac{\hat{p}_{i,j}^{(1-\beta_i)}}{\sum_{k=1}^{K} \hat{p}_{i,k}^{(1-\beta_i)}}$$

Here, βi is the probabability assigned by biased model to the gold label for that particular example

If βi is close to 1, the resulting probability distribution is become 1/3 for all.
If βi is close to 0, the probability distribution is unchanged.

Results

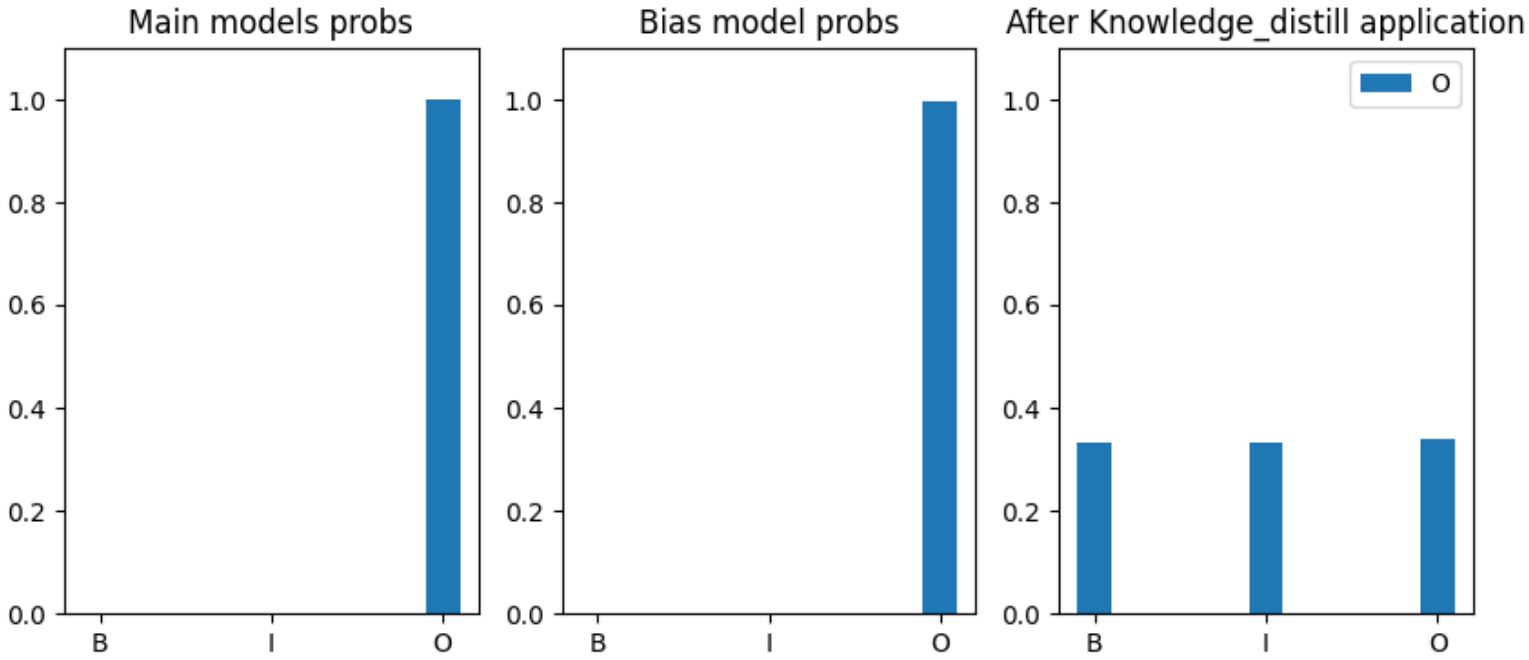
NCBI-disease						
Model (early stopped)	Precision	Recall	F1-score	Mem	Syn	Con
Biobert	83.7	89.7	86.6	94.6	77.5	85.8
BiLSTM bias	23.4	87.7	37.0	91.9	80.6	80.2
Prior Prob	79.5	83.6	81.5	87.9	77.5	74.7

BC5CDR						
Model (early stopped)	Precision	Recall	F1-score	Mem	Syn	Con
Biobert	82.9	89.5	86.1	93.9	81.3	84.4
BiLSTM bias	48.6	86.9	62.3	91.3	79.2	80.7
Prior Prob	69.7	85.3	76.8	92.1	73.6	75.5

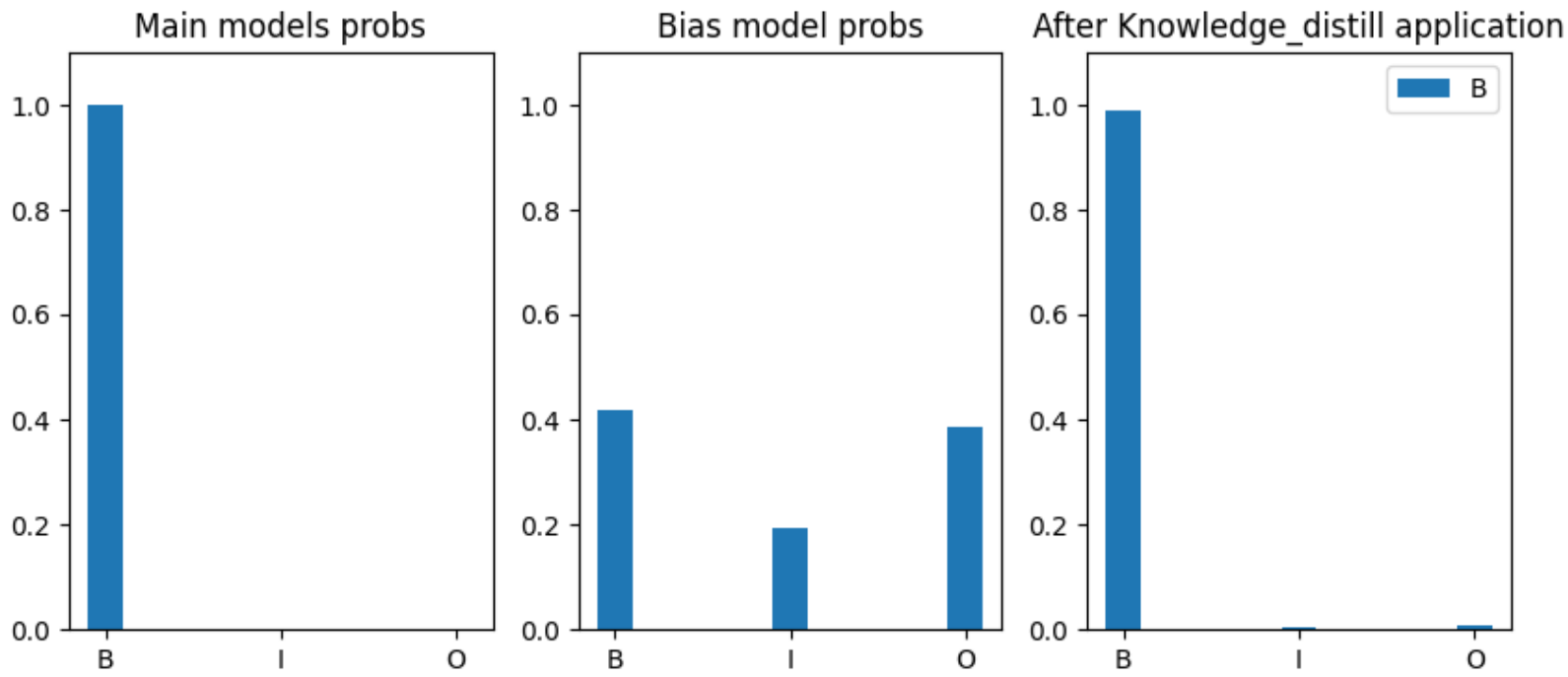
MedMentions						
Model (early stopped)	Precision	Recall	F1-score	Mem	Syn	Con
Biobert	65.6	68.9	67.2	70.4	65.4	65.6
BiLSTM bias	56.8	69.6	62.5	71.2	65.5	65.3
Prior Prob	64.5	69.5	66.9	70.9	66.2	65.9

Why Confidence Regularization works poorly for BC5CDR and NCBI-disease, but not that bad for MedMentions?

Effect of scaling



When bias model predicts correct label with high confidence



When bias model predicts correct label with low confidence

Our **bias models are too strong**, they are predicting many of the tokens with high confidence.

Dataset	Bias Model	≥ 0.85	≥ 0.90	≥ 0.95
BC5CDR	Prior Probability	88%	87%	85%
	BiLSTM	93%	90%	86%
NCBI-Disease	Prior Probability	84%	83%	81%
	BiLSTM	90%	89%	88%
MedMentions	Prior Probability	55%	47%	40%
	BiLSTM	64%	57%	47%

Table: Percentages of examples having assigned probability to gold label higher than a particular threshold

Debiasing using Biased Committee

This method borrows some parts from this paper [4].

Uses a committee of classifiers to weight an examples importance

Biased committee – Consists of 96 MLPs (2 layers, ReLU activation)

Two step process:

- Training of committee
  - Non-overlapping subsets of data for each classifiers
- Weighted training of BioBERT
  - Used two types of weighting functions:
    - Linear weighting
    - Non-linear weighting

[4] Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. **Learning debiased classifier with biased committee.** Advances in Neural Information Processing Systems, 2022

Linear Weighting

$$w(x) = 1 - \frac{\sum_{l=1}^{L} 1(f(x) = y)}{L}$$

Non-linear Weighting

$$w(x) = \frac{1}{\sum_{l=1}^{L} 1(f(x) = y) + \alpha}$$

- α is the hyperparameter, we used its value as 1

- If more the classifiers in committee predict correctly, the weight goes toward 0.
- Also uses a hyperparameter called cutoff number. If the number of classifiers predicting correclty > cutoff_num, assign 1 as weight.

Results from bias committee

NCBI-disease						
Model (early stopped)	Precision	Recall	F1-score	Mem	Syn	Con
Biobert	83.7	89.7	86.6	94.6	77.5	85.8
Linear weighting	87.0	90.3	88.6	95.5	79.1	84.6
Non-linear weighting	87.0	91.1	89.0	94.3	84.3	87.7

BC5CDR						
Model (early stopped)	Precision	Recall	F1-score	Mem	Syn	Con
Biobert	82.9	89.5	86.1	93.9	81.3	84.4
Linear weighting	84.7	89.6	87.1	95.7	78.6	82.3
Non-linear weighting	83.9	90.0	86.9	95.2	80.7	83.5

MedMentions						
Model (early stopped)	Precision	Recall	F1-score	Mem	Syn	Con
Biobert	65.6	68.9	67.2	70.4	65.4	65.6
Linear weighting	64.6	68.5	66.5	70.8	63.0	63.1
Non-linear weighting	64.6	70.5	67.4	73.4	63.4	63.6

Relaxed Matching

Issues with predictions

- Issues with spans
- Adjective non-uniformly included in mentions

Rules of relaxed matching:

- B and I treated as same tag
- A difference in tags when a token is adjective is ignored

Relaxed Matching metric						
Dataset	Precision	Recall	F1-score	Mem	Syn	Con
BC5CDR	86.1	94.2	90.0	98.7	88.2	90.9
NCBI-disease	95.1	95.6	95.4	98.7	92.6	92.0
MedMentions	86.0	90.3	88.1	89.0	91.9	92.2

Issues with dataset annotation

- Dataset annotation has some non-uniformities in labelling adjective for mentions.
- It has some omissions, where the tokens must be part of mention.

Issues in dataset annotation. Grey stands for O-tag, Orange for B-tag, and Green for I-tag.

Original Dataset	Model Prediction
stess ulcers	stress ulcers
AL amyloidosis	AL amyloidosis
depressive -like behaviour	depressive -like behaviour
convulsive seizure	convulsive seizure
methamphetamine dependence	methamphetamine dependence
sporadic Alzheimer's disease (sAD)	sporadic Alzheimer's disease (sAD)

Thank You