Project 3- N2C2 2018 Track 2: Named Entity Extraction

Soujanya Ranganatha Bhat, Varun Chaudhary
{sranga16, vchaudh7}@asu.edu

1 Task Review

The N2C2 2018 track 2 dataset comprises raw text of narrative discharge summaries(1), drawn from the MIMIC-III clinical care database. The training data is of the form **<ID, Tag, Position, Phrase>**. The given data contained two parts- training(further split into validation) and test sets. The task can be summarised as identifying the following entities from the discharge summaries- **Drug, Reason, ADE(Adverse Drug Event)**.

An example of the data can be found below:

- **Discharge summary text**: He may also have recurrent seizures.
- **Annotated data**: <T1, Reason, 18 36, recurrent seizures >

The original tag distribution can be found in table 1.

Type	Full	Training	Test
Drug	26800	16225	10575
Reason	6400	3855	2545
ADE	1584	959	625
Total	34784	21039	13745

Table 1: Data split and class distribution

2 Method

2.1 Proposed Method Description

Our task pipeline broadly consists of the following parts-data preparation, training biomedical BERT for NER downstream task and model performance analysis on the test set.

During the pre-processing phase, the discharge summary was tokenized(using nltk) into BIO-tag format and the processed data was saved as <PatientID, SentenceID, Token, Tag >in a CSV file.

For our model, we decided to replace the previously used softmax layer with a Conditional Random Fields(CRF) layer(2). So, BioDischargeSummaryBERT's(BDSBERT) encoded token sequence(with hidden dimension) were projected to the tag space by a linear classification model. The output scores were fed to the CRF layer, which has a tag transition matrix of transitions $(K+2)*(K+2)$, where K is the number of tags and matrix includes two additional states such as start and end of sequence. The value in each cell represents the score of transitioning from one tag to another. The model is trained to maximise the log-probability of the correct tag sequence. During evaluation, the most likely sequence is obtained by Viterbi decoding.

2.2 Innovation

We referred to the the existing state-of-the-art BiLSTM-CRF model for motivation. We leveraged BERT's generally superior word embeddings to its BiLSTM alternative. While our BSDBERT model performed decently, it was drastically inferior to the SOTA for ADE entities. So, we decided to experiment with the following techniques:

- Under-sampling O tags or Over-sampling underrepresented classes(negligible gain in performance).
- Weighted cross entropy loss(minute gain in performance).
- **Replaced the softmax layer with a CRF layer(explained in the method description), to enforce sequential classification(significant gain in performance).**

3 Experiment Setup

3.1 Baseline Method

We make use of two baselines for comparison. For the purpose of having a BERT baseline, we updated to our developed vanilla BDSBERT model. The second baseline is reported as part of the N2C2 2018 Task 2 challenge which is a BiLSTM-CRF model variant(with pre-processing techniques). The available scores are reported in the table 2.

	Drug	Reason	ADE
BDSBERT	P- 0.9 R- 0.93 $F1_l$- 0.92	P- 0.48 R- 0.7 $F1_l$- 0.57	P- 0.17 R- 0.52 $F1_l$- 0.26
BiLSTM-CRF	$F1_l$- 0.58	$F1_l$- 0.72	$F1_l$- 0.96

Table 2: Baselines

3.2 Dataset Separation

We utilise three sets after splitting train $\rightarrow$ train & val, resulting in a 65-10-25 ratio split. The data folders provided were used as following- train(track2-training_data_2), validation(track2-training_data_3) and test(gold-standard-test-data). The training dataset is used to aid the model in learning the required BIO-tagging for concept extraction, while validation set keeps a check on over-fitting. Finally, the model performance is evaluated on the test set. The class distribution across each split can be found in the table 3.

3.3 Evaluation Metrics

We analysed the **confusion matrix** to gain insights and examine the model performance. For model evaluation, we used micro-avg, macro-avg and weighted-avg F1

Analysis	Category	Examples	Actual	Predicted	Explanation
Improvement	Better learnt Reason	recurrent disease	Reason	Reason(now) Unknown(earlier)	Reason are identified more often (even with only 18% entities as Reasons)
	Better learnt ADE	intractable diarrhea	ADE	ADE(now) Unknown(earlier)	ADEs are identified more often (even with only 4% entities as ADEs)
	Breaks in terms(hyphens, slashes)	beta-lactamase inhibitor combinations	Drug	Drug(now) Unknwon till hypen, then Drug(earlier)	Model now predicts hyphenated words much better
Error	Tag Overlap	vincristine vincristine toxic polyneuropathy	Drug ADE	Drug Drug	Same phrase consists of two tags, where the model identifies only one
	Abbreviations	SVC syndrome	Reason	Unknown	Superior vena cava (SVC) syndrome and other abbreviations is not identified
	Ambiguity	allergic reaction	Reason	ADE	Ambiguous use of Reasons and ADEs, without a distinct difference
	Complicated terms	mediastinal lymph nodes measuring up to 9 mm diffuse erythema/ulceration, in esophagus, stomach, dupdenum	Reason Reason	Unknown Unknown	Entity spans over multiple complicated words

Figure 1: Analysis(Improvements and Errors)

Type	Training	Validation	Test
Drug	14303	2045	10569
Reason	3258	445	2519
ADE	813	103	604
Total	18374	2593	13692

Table 3: Class distribution(pre-processed data)

scores, both lenient($F1_l$) and strict($F1_s$) along with precision and recall on entity level. We compared these metrics against our baselines to gain idea of our model's standing.

4 Result

4.1 Results of the Test set

The model performance based on the previously mentioned evaluation metrics is shown in the Table 4 for entity-level analysis and Table 5 for overall F1 scores. We also compare the various models in Table 6.

	P	R	$F1_l$	$F1_s$
Drug	0.95	0.96	0.96	0.92
Reason	0.76	0.71	0.73	0.60
ADE	0.73	0.55	0.63	0.50

Table 4: Entity-level Results($F1_l$- lenient, $F1_s$- strict)

Type	Score
Micro $F1_l$	0.90
Macro $F1_l$	0.77
Weighted $F1_l$	0.90
Micro $F1_s$	0.85
Macro $F1_s$	0.68
Weighted $F1_s$	0.85

Table 5: Overall F1 scores

The Improvements in the scores (from the midterm) can be attributed to the following enhancements:

- Analyzed and determined the correct input sequence length so the BERT word pieces do not lead to information truncation on exceeding maximum sequence length.
- Efficiently filtered tag overlaps(multiple entities in the same span) by prioritising entity span length to include more ADE and Reason entities over the largely populated Drug entities.
- Increased the number of training epochs to 100 for baseline BDSBERT to achieve better score **however**, with BDSBERT-CRF, we achieved better F1 scores at the 13^{th} training epoch as CRF layer can leverage its better structures predictions via the state transition matrix in contrast to a simple softmax layer.

Models	$Drug_{F1_l}$	$Reason_{F1_l}$	ADE_{F1_l}
BDSBERT(midterm)	0.68	0.21	0.06
BDSBERT(baseline)	0.92	0.57	0.26
BiLSTM-CRF(SOTA)	**0.96**	0.72	0.58
BDSBERT-CRF(ours)	**0.96**	**0.73**	**0.63**

Table 6: Model Comparison

4.2 Error Analysis

We analysed the model performance especially in the areas of significant improvement such as ADE and Reason entities:

- The confusion matrix in Table 7 highlights that the model still cannot classify part of the Reason and especially ADE entities, although it is significantly better than the previously reported results.

	Drug	Reason	ADE	O
Drug	**10153**	4	1	411
Reason	13	**1786**	23	697
ADE	2	40	**332**	230

Table 7: Confusion matrix

- The error categories along with areas of improvement in comparison to the previously reported model can be found in Figure 1.

References

[1] *Henry, S., Buchan, K., Filannino, M., Stubbs, A. and Uzuner, O., 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association, 27(1), pp.3-12.*

[2] *Souza, F., Nogueira, R. and Lotufo, R., 2019. Portuguese named entity recognition using BERT-CRF. arXiv preprint arXiv:1909.10649.*