Izvješće o eksperimentima

Eksperimenti su dokumentirani tako da je naziv foldera u koji se spremaju log-ovi (ujedno i naziv modela) sljedeći:

→ D1/D2 - označava na kojem je datasetu rađen (D1-NCBI-disease, D2-BC5CDR)
→ B/E - označava koji je word embedding korišten (B-bioBERT, E-bioEMLo)
→ C/_ - označava koristi li se CNN char embedding (C-da, _-ne)
→ L/G - označava vrstu ćelije RNN-a (L-LSTM, G-GRU)
→ <broj> - broj koji označava konfiguraciju preostalih hiperparametara (svaki - eksperiment u config.log datoteci sadrži sve parametre s kojima je pokrenut)
→ A – A označava da se koristi attention

Primjeri:
D1_B_C_L_1 - označava eksperiment gdje je model treniran nad NCBI-disease datasetu, uz bioBERT embedding, koristeći char CNN i uz LSTM ćeliju, riječ je o prvoj kombinaciji hiperparametara i ne koristi se attention.
D2_E___G_1 - označava eksperiment gdje je model treniran nad BC5CDR datasetu, uz bioELMo embedding, ne koristeći char CNN i uz GRU ćeliju, riječ je o prvoj kombinaciji hiperparametara i NE koristi se attention.

D2_E___G_1_A - označava eksperiment gdje je model treniran nad BC5CDR datasetu, uz bioELMo embedding, ne koristeći char CNN i uz GRU ćeliju, riječ je o prvoj kombinaciji hiperparametara i koristi se attention.

Implementacijski detalji*:*

- Early stopping je postavljen tako da se gleda f1-mjera (strict) na validacijskom skupu podataka.
- Dodan je lr_scheduler – ReduceLROnPlateau s parametrima (`mode='max', factor=0.5, patience=2`) koji isto gleda f1 (strict) na validacijskom skupu podataka i ako se dvije epohe za redom vrijednost ne poveća, smanjuje lr za faktor 0.5. (*Napomena: probleme s ekplodirajućim loss-om koje ovo rješava sam uočila tek nakon implementacije attentiona pa je lr_sceduler prisutan samo u *_A modelima*)
- Za sada je implementiran multi-head attention layer koji na ulazu ima skriveni sloj RNN-a (uzeta je Pytorch implementacija sloja) – 2 rada koja su koristila ovu vrstu attention-a i poslužila kao inspiracija: https://ieeexplore.ieee.org/document/8798611 i https://arxiv.org/pdf/2002.00735. Još je stavljena konkatenacija izlaza multi-head attention sloja i skrivenog sloja RNN-a (isprobano je sa i bez

i malo bolji rezultat je kad se napravi konkatenacija pa je ovako trenutno implementirano) – to su *_A modeli.

Postavke za konfiguraciju 1:

hidden_size: 512
num_layers: 1
dropout: 0.3
learning_rate: 0.001
batch_size: 32
optimizer: "adam"
epochs: 300
max_length: 256
max_grad_norm: 5.0
early_stopping: 10
cnn_vocab: "abcdefghijklmnopqrstuvwxyz0123456789-,;.!?:'\"/\\|_@#$%^&*~`+-=<>()[]{}"
cnn_max_word_len: 20
cnn_embedding_dim: 256
feature_size: 256
att_num_of_heads: 16

U tablici koja slijedi, prikazane su dobivene f1-mjere po skupovima podataka. Za bc5cdr prikazana je odvojena f1-mjera po entitetima (chem-kemikalija, dis-bolest) te mikro usrednjena f1-mjera na cijelom skupu podataka. Također, za oba skupa podataka, prikazan je i defaul i strict izračun. (Napomena: strict način rada osigurava da se predikcije entiteta računaju kao točne isključivo ako potpuno odgovaraju stvarnim granicama entiteta i njegovom tipu).

MODEL	BC5CDR-CHEM	BC5CDR-DIS	BC5CDR	NCBI-DIS	BC5CDR-CHEM (strict)	BC5CDR-DIS (strict)	BC5CDR (strict)	NCBI-DIS (strict)
B_C_G_1	0.84	0.78	0.81	0.76	0.38	0.74	0.64	0.73
B___G_1	0.84	0.76	0.81	0.77	0.44	0.74	0.67	0.73
B___L_1	0.85	0.77	0.81	0.78	0.41	0.74	0.66	0.75
B_C_L_1	0.85	0.78	0.82	0.79	0.46	0.75	0.68	0.75
B___L_1_A	0.9	0.8	0.85	0.82	0.58	0.76	0.73	0.78
B_C_L_1_A	0.89	0.81	0.85	0.82	0.6	0.76	0.73	0.78
B___G_1_A	0.9	0.8	0.86	0.77	0.56	0.76	0.72	0.75
B_C_G_1_A	0.9	0.82	0.86	0.82	0.61	0.78	0.74	0.79
E___G_1	0.93	0.82	0.88	0.78	0.72	0.77	0.76	0.75
E_C_G_1	0.92	0.82	0.88	0.79	0.71	0.77	0.76	0.76
E___L_1	0.93	0.82	0.88	0.8	0.79	0.76	0.77	0.76
E_C_L_1	0.93	0.82	0.88	0.82	0.79	0.77	0.77	0.79
E_C_L_1_A	0.93	0.84	0.89	0.84	0.79	0.78	0.79	0.79
E___G_1_A	0.94	0.84	0.89	0.84	0.8	0.78	0.79	0.79
E___L_1_A	0.94	0.84	0.89	0.84	0.81	0.78	0.79	0.8
E_C_G_1_A	0.94	0.84	0.89	0.85	0.81	0.79	0.79	0.8

Analiza i zaključci:

-> Modeli koji koriste bioELMo imaju bolje performanse (možda jer bioBERT nije large model, tu bi se to moglo još isprobati)
-> Dodavanje ovako definiranog char CNN embeddinga ne pomaže nešto puno
-> Dodavanje ovako definiranog attention sloja isto ne pomaže puno
-> Strict f1-mjere (očekivano) niže
-> Trenutno najbolji modeli - E_C_G_1_A

Plan za dalje:

-> Proučiti i implementirati drugu vrstu attentiona (lokalna - na razini tokena, više bazirana na susjedstvo)
-> Proučiti i implementirati dice loss (*pitanje: kombinirati nekako s CRF loss-om i kako? ili koristiti zasebno, tj. umjesto CRF loss-a*)
-> Proučiti i implementirati multitask