به نام خدا

تارا قشلاقی- 99443009

لینک

لینک پیاده‌سازی finetuning و نتایج (بخش ***Test model***) در لینک موجود است.

در فاین تیون کردن از مدل parsber v2 استفاده شده است که در حین مقدار دهی کردن tokenizer، config و model از

```
model_name_or_path = 'HooshvareLab/bert-fa-base-uncased'
```

استفاده کرده‌ایم.

در ابتدا داده های موجود در فایل txt. را استخراج می‌کنیم که با توجه به پیکره‌ی PEYMA بعد از هر خط خالی، جمله‌ی جدیدی شروع خواهد شد و تگ و کلمه نیز با | از یکدیگر جدا شده اند پس با استخراج جملات و کلمه ها و تگ ها، عملیات پیش‌پردازش را انجام می‌دهیم تا داده آماده‌ی ورودی دادن به مدل برت شود.

به این صورت که بعد از فراخوانی tokenizer مربوطه روی هر کلمه که خود ممکن است شامل چند توکن شود، تگ آن کلمه را به تعداد توکن هایی که حاصل از tokenize شدن هستند تکرار می کنیم و در لیست مربوط به تگ ها متناظر با جمله، اضافه می شود سپس توکن ها را به عدد تبدیل می کنیم (قابل ورودی دادن به مدل)

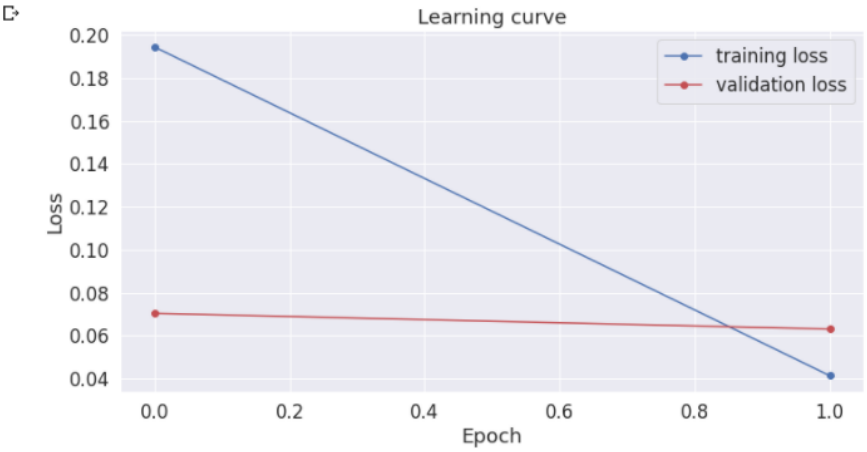
همچنین لیبل ها را نیز به عدد تبدیل می کنیم (از قبل به هذ تگ عدد خاصی نسبت می دهیم)

سپس برای عملیات padding، ابتدا maxlenرا به دست می‌آوریم ولی عددی حدود ۴۰۰ به دست آمد که طبق این باید maxlen را برابر با ۵۱۲ قرار داد و همین باعث می شد به دلیل حافظه، کد ارور بدهد. بنابراین از عدد رایج ۳۲ در بیشتر کدها استفاده کردیم.

(دلیل عدد ۴۰۰ برای تعداد توکن های یک عنصر به این دلیل بود که در فایل داده جمله های آن از هم جدا نشده بود بنابراین از آن چشم‌پوشی شد)

سپس بعد از انجام دادن attention mask و چشم پوشی از مقادیر صفر حاصل از padding برای آن، و عملیات تقسیم داده، و تبدیل داده ها به tensor، حال ورودی مدل آماده است.

مدل pretrained ، در دو ایپاک train شد و نتایج حاصل از loss validation همانطور که در لینک کلب هم آمده به صورت زیر است:



هم چنین مدل به دست آمده را بر روی داده های تست نیز بررسی کردیم که در لینک کلب در بخش Test model آمده است و نتایج به صورت زیر می‌باشد:

در حالتی که بین اینکه تگ اول یا در ادامه entity آمده تفاوت قائل شدیم؛

	precision	recall	f1-score	support
I_DAT	0.831	0.948	0.886	249
B_DAT	0.851	0.832	0.841	220
B_PER	0.956	0.954	0.955	497
B_MON	0.800	0.923	0.857	26
I_ORG	0.899	0.903	0.901	1149
B_LOC	0.914	0.947	0.930	607
I_MON	0.953	0.924	0.938	66
I_PCT	0.927	0.905	0.916	42
B_ORG	0.899	0.874	0.887	716
O	0.996	0.992	0.994	32507
B_TIM	0.889	0.727	0.800	22
B_PCT	0.904	0.940	0.922	50
I_PER	0.933	0.964	0.949	363
I_TIM	0.821	0.958	0.885	24
I_LOC	0.718	0.900	0.798	229
PAD	0.000	0.000	0.000	0
micro avg	0.983	0.983	0.983	36767
macro avg	0.831	0.856	0.841	36767
weighted avg	0.984	0.983	0.984	36767

در حالتی که تگ را معیار قرار دادیم نه اینکه اول آمده یا در ادامه‌ی entity:

	precision	recall	f1-score	support
_DAT	0.75	0.78	0.76	220
_LOC	0.87	0.92	0.89	607
_MON	0.65	0.77	0.70	26
_ORG	0.83	0.83	0.83	716
_PCT	0.81	0.88	0.85	50
_PER	0.91	0.92	0.91	497
_TIM	0.75	0.68	0.71	22
micro avg	0.85	0.87	0.86	2138
macro avg	0.79	0.82	0.81	2138
weighted avg	0.85	0.87	0.86	2138