به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس پردازش زبان طبیعی

تمرین شماره پنج

نام و نام خانوادگی:سید مهدی موسوی

شماره دانشجویی :810102264

خرداد 1403

فهرست سوالات

مقدمه

در این تمرین سعی داریم که با استفاده از دادگان در دست مدلی طراحی بکنیم که جملات انگلیسی را دریافت و ترجمه فارسی آن ها را به ما بدهد. این مدلسازی با استفاده از لایه های LSTM و attention انجام میشود و در نهایت با استفاده از معیار های comet و bleu آن ها را ارزیابی میکنیم.

دادگان

در اولین قسمت از مراحل انجام پروژه دادگان در درست را بررسی و در ادامه در صورت نیاز اصلاح میکنیم. دادگان در دست مجموعه بزرگی از دادگان است که شامل جملات انگلیسی و جملات متناظر ترجمه آنان است. تعداد کل جملات در هر زبان 1021597 عدد است.

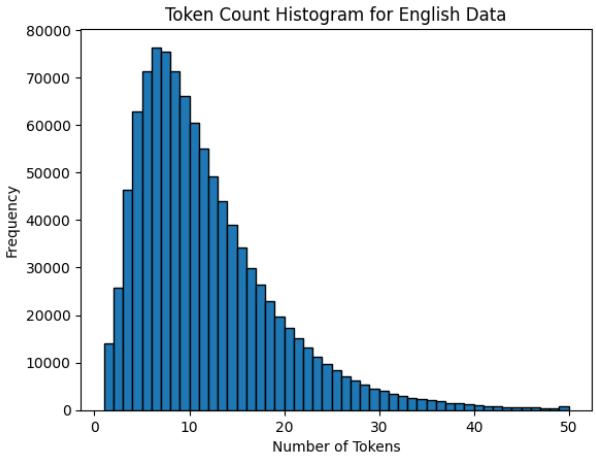
در ادامه کار سه خط اول دادگان را برای هر دو زبان چاپ میکنیم. جملات انگلیسی به صورت زیر است :

- The story which follows was first written out in Paris during the Peace Conference
- from notes jotted daily on the march, strengthened by some reports sent to my chiefs in Cairo.
- Afterwards, in the autumn of 1919, this first draft and some of the notes were lost.

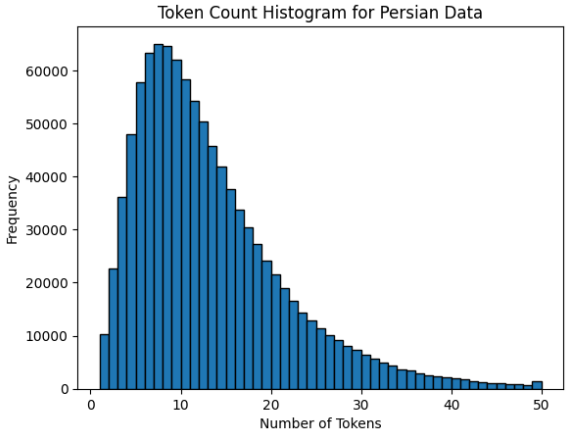
و معادل آن ها در فارسی :

- داستانی که از نظر شما می‌گذرد، ابتدا ضمن کنفرانس صلح پاریس از روی یادداشت‌هائی که به طور روزانه در حال خدمت در صف برداشته شده بودند .
- و از روی گزارشاتی که برای رؤسای من در قاهره ارسال گردیده بودند نوشته شد .
- بعدا در پائیز سال 1919، این نوشته اولیه و بعضی از یادداشت‌ها، مفقود شدند.

همچنین برای توزیع اندازه جملات نمودار های 1-1 و 1-2 را خواهیم داشت .



نمودار ۱-۱ : توزیع تعداد توکن های جملات انگلیسی



نمودار ۲-۱ : توزیع تعداد توکن های جملات فارسی

مطابق خواسته سوال و برای همگن شدن دادگان سعی میکنیم که جمله هایی که در داده فارسی کمتر از 10 توکن یا بیشتر از 50 توکن دارند را از دادگان حذف کنیم. بعد از این کار تعداد جملات به عدد 585266 رسید که به معنای تقریبا نصف شدن دادگان است.

در آخرین مرحله هم مطابق موارد مطرح شده در سوال 50000 سطر برای آموزش, 5000 سطر برای ارزیابی و 10000 سطر برای آزمون جدا شدند و در فایل های با نام های متناظر و متناسب ذخیره شدند.

پیش پردازش

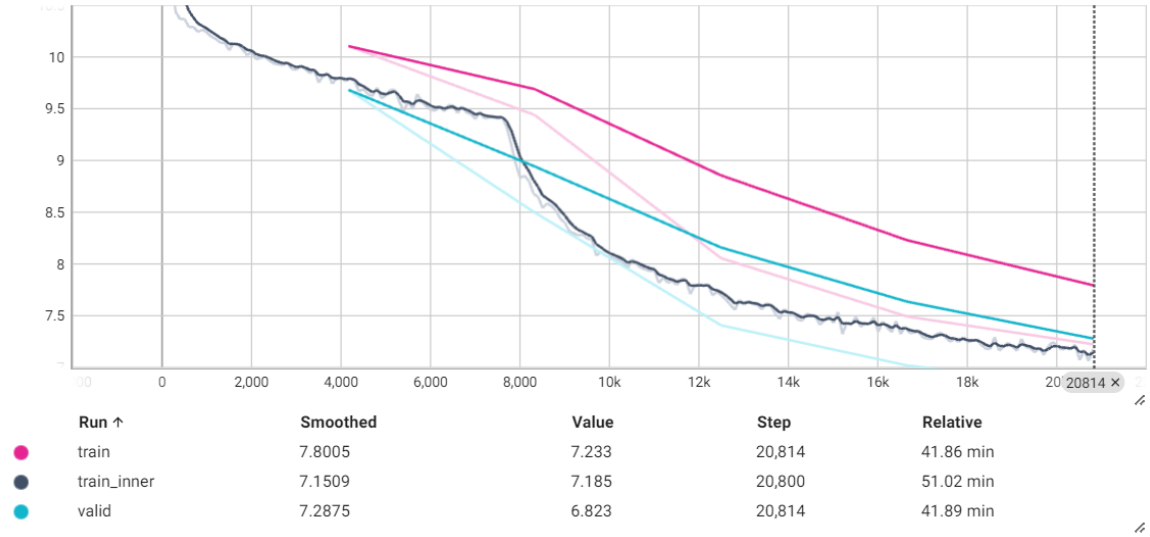
در دومین مرحله سعی میکنیم که پیش پردازش های لازم را روی داده انجام بدهیم تا بتوانیم دادگان را مطابق ورودی استاندارد مدل هایی که در ادامه استفاده میشوند بکنیم. تابع استفاده شده در این قسمت خروجی های لازم برای مدل را در فرمت باینری ذخیره میکند. همچنین پارامتر های مشخص شده در صورت سوال برای این هستند که اندازه دیکشنری انگلیسی و فارسی 10000 باشند.

هدف اصلی این بخش این است که فایل های مورد نیاز برای آموزش به فرمت باینری ذخیره شود. با استفاده از این روش مقدار زیادی از سر بار سیستمی کم میشود و به تبع آن مدل سریعتر میشود. به همین ترتیب استفاده از حافظه هم با این روش کاهش می یابد و در نهایت اینکه به کمک این تابع ورودی های مدل فرمت یکسان خواهند بود. همچنین در این فرمت پردازش موازی هم راحتتر میشود.

در نهایت خروجی ما شامل دو فایل دیکشنری برای انگلیسی و فارسی و همچنین فایل های باینری شده و ایندکس میشود.

آموزش مدل بر پایه LSTM

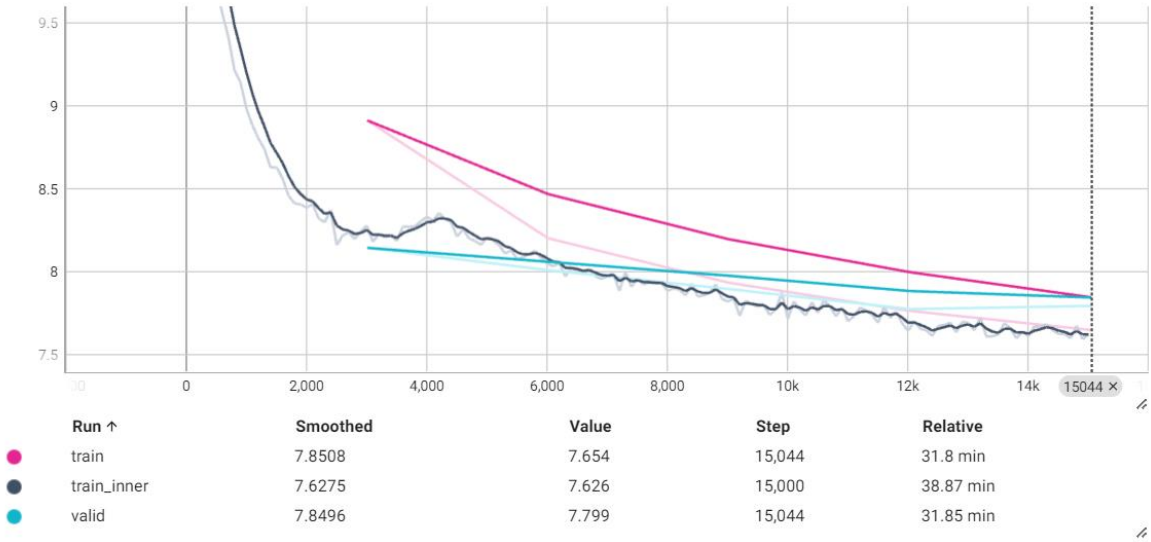
در اولین قسمت سعی میکنیم که مدل را به کمک لایه های LSTM آموزش دهیم. از جمله پارامتر های استفاده شده در حین آموزش میتوان به max tokens اشاره کرد که کاری که این پارامتر انجام میدهد این است که حداکثر تعداد توکن های یک batch را مشخص میکند و این پارامتر به خصوص چون طول جملات ما متغیر است در این کاربرد مهم است. پارامتر دیگری که مطرح است batch size است که تعداد جملات موجود در هر batch است.مقادیر استفاده شده برای این دو پارامتر 4096 و 128 است به این معنا که به طور میانگین اگر کمتر از 32 باشد میتوانیم تمام جملات را در یک batch جا بدهیم و در غیر این صورت مجبور هستیم تعداد کمتری از جملات را در نظر بگیریم. در نهایت نمودار تغییرات تابع هزینه برای این مدل در نمودار 3-1 آمده است.



نمودار 3-1: نمودار تغییرات تابع هزینه برای مدل LSTM در حین آموزش

آموزش مدل بر پایه attention

در مرحله بعدی سعی میکنیم که از مدل های مبتنی بر attention استفاده کنیم.دادگان استفاده شده در این قسمت هم دقیقا مشابه قسمت قبل است. در نهایت برای این قسمت هم نمودار تغییرات هزینه به دست آوردیم که در نمودار 2-3 دیده میشود.



Run ↑	Smoothed	Value	Step	Relative
train	7.8508	7.654	15,044	31.8 min
train_inner	7.6275	7.626	15,000	38.87 min
valid	7.8496	7.799	15,044	31.85 min

نمودار 2-3: نمودار تغییرات تابع هزینه برای مدل attention در حین آموزش

نتیجه گیری و ارزیابی

در آخرین قسمت سعی میکنیم که مدل به دست آمده را به کمک معیار های در دست ارزیابی کنیم. برای مدل آموزش داده شده به کمک LSTM مقدار معیار BLUE برای داده آزمون 5.11 و این معیار برای مدل مبتنی بر attention برابر 0.4 شد.

در وهله بعد سعی میکنیم که comet score را برای دو مدل مطرح شده محاسبه کنیم. Crosslingual Optimized Metric for Evaluation of Translation یا COMET یک معیار مطرح شده برای ترجمه ماشینی است که کاری که انجام میدهد این است که برای جملات مبدا و مقصد embedding تشکیل میدهد و در نهایت سعی میکند که فاصله آنها را تبدیل به یک عدد و معیار کند که این کار را به کمک یک تابع رگرسور انجام میدهد که البته مدل و فرمول دقیق امتیاز دهی بسته به مدل انتخاب شده متفاوت است.

از مزایای این معیار به نسبت معیار قبلی میتوان به این موارد اشاره کرد که در این روش مثل BLUE صرفا به n-gram ها توجه نداریم و به معنای جملات هم توجه داریم و اینکه مفهوم و معنای کلی هم در بازنمایی تاثیر دارد که در نهایت در امتیاز تاثیر میگذارد.

```
                                                  _she _is _always _upon _the _g ad  S-7267
                                          _از _جمله _این _که _همیشه _مشغول _گردش _و _تفریح _است  T-7267
                                  _او _همیشه _در _این _باره _به _سر _می _برد .     -2.1855030059814453  H-7267
                                  _او _همیشه _در _این _باره _به _سر _می _برد .     -2.1855030059814453  D-7267
-2.6263 -0.9248 -2.5525 -3.4499 -3.6578 -3.1524 -3.4433 -0.7446 -0.9703 -2.4044 -0.1144  P-7267
```

نمودار 3-3 : یک نمونه از خروجی مدل

در نهایت امتیاز comet برای LSTM 0.523 و برای attention برابر 0.438 شد که مشابه امتیاز BLUE در این قسمت هم مدل LSTM عملکرد بهتری داشت که این موضوع به نظر به دلیل کم بودن منابع است.