به نام آنکه آموخت انسان را آنچه نمی دانست

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس پردازش زبان‌های طبیعی

CA6

NMT

En2Fa Translation + Fa2En Transliteration

behzad.shayegh@ut.ac.ur

بهزاد شایق بروجنی

81 01 96 678

اردیبهشت ۱۳۹۹

فهرست

مقدمه

با سلام. در این تمرین قصد داریم با کتابخانه‌ی open nmt آشنا شده و دو مسئله‌ی ترجمه‌ی انگلیسی به فارسی و همچنین نویسه گردانی فارسی به انگلیسی را با استفاده از آن حل کنیم. جزئیات پیاده سازی (به دلیل حجم زیاد) تشریح نخواهد شد؛ برای اطلاع از این جزئیات می‌توانید به کدهای ضمیمه شده مراجعه کنید.

بخش اول) ترجمه انگلیسی به فارسی

- دادگان

دادگان این بخش، سه دسته دادگان train, test, dev می‌باشد که هر دسته شامل دو فایل متنی موازی از جملات انگلیسی و فارسی معادل است. مجموعه test دارای ۴ نسخه مرجع ترجمه فارسی برای ارزیابی انعطاف‌پذیرتر است.

زیر بخش اول) بدون استفاده از bpe

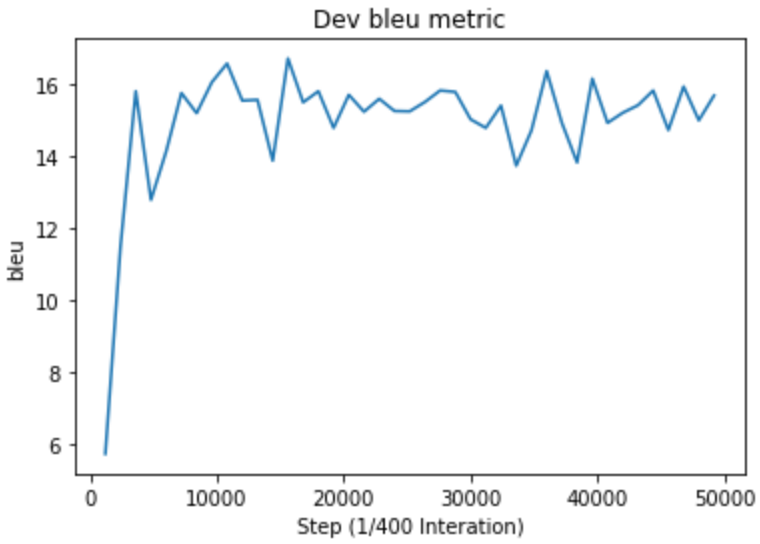
- الف) معیار bleu بر روی دادگان ارزیابی - منبع : NLP_CA6_Part1_default.ipynb

پس از ۵۰۰ iteration آموزش شبکه، معیار bleu برای دادگان تست و چهار مرجع ترجمه به شکل زیر بدست آمد (به ترتیب مراجع) :

```
BLEU = 17.12, 58.9/27.1/13.1/6.0 (BP=0.910, ratio=0.914, hyp_len=2461, ref_len=2692)
BLEU = 18.38, 59.9/28.4/14.4/7.0 (BP=0.903, ratio=0.907, hyp_len=2461, ref_len=2713)
BLEU = 12.50, 52.3/20.6/9.1/3.7 (BP=0.907, ratio=0.911, hyp_len=2461, ref_len=2700)
BLEU = 13.53, 51.5/21.1/9.2/4.0 (BP=0.955, ratio=0.956, hyp_len=2461, ref_len=2575)
```

- ب) نمودار bleu بر روی دادگان توسعه - منبع : NLP_CA6_Part1_default.ipynb

نمودار مقادیر bleu بر روی دادگان توسعه، بر حسب تعداد iteration آموزش به صورت زیر بدست آمد (دقت این نمودار ۳ می‌باشد به این معنا که معیار blue برای هر ۳ iteration محاسبه شده است. همچنین این نمودار بر حسب گام آموزشی رسم شده که هر ۴۰۰ گام معادل یک iteration بود.) :



همانطور که مشاهده می‌کنید، این معیار خیلی سریع (تقریبا بعد از گام 8000 که معادل iteration بستم است) به همگرایی می‌رسد.

- پ) خطایابی دستی - منبع : NLP_CA6_Part1_default.ipynb

با کاوش در پاسخ‌های مدل به دادگان تست (بعد از iteration آخر)، متوجه انواع خطا شدیم که سه نمونه را در ادامه بررسی می‌کنیم :

۱. طول جمله

به دو نمونه زیر توجه کنید :

نمونه شماره 223 :

during a flight I always get sick . I would like to take the train .

ترجمه های مرجع :

0.من همیشه در طول پرواز حالم بد میشود . دوست دارم سوار قطار شوم .
1.من همیشه درطول یک پرواز مریض میشوم . من مایلم قطار بگیرم .
2.من همیشه در طول مدت پرواز مریض میشوم . من ترجیح میدهم که قطار بگیرم .
3.درطی پرواز من همیشه مریض میشوم . من دوست دارم قطار بگیرم .

ترجمه ماشین :

پرواز دارم من همیشه . دوست دارم از .

==========

نمونه شماره 177 :

that is fine . when does the plane leave Hanover ?

ترجمه های مرجع :

0.این خوب است . چه وقت هواپیما هانوور را ترک میکند ؟
1.این خوب است . کی هواپیما هانوفر را ترک میکند ؟

2.این خوب است . هواپیما کی هانوور را ترک میکند ؟

3.آن خوب است . چهزمانی هواپیما هانوور را ترک میکند ؟

ترجمه ماشین :

بله . چه زمانی است که دارد ؟

==========

همانطور که واضح است، مدل ما سعی بر کوتاه کردن جملات خروجی داشته و همین باعث کاسته شدن از کیفیت ترجمه‌ی آن شده است. دلیل پیش‌آمدن این موضوع، بالا بودن دقت ترجمه درصورت کمتر بودن تعداد لغات ترجمه است. راه‌حلی که ممکن است این مشکل را برطرف کند، زیاد کردن پنالتی طول کوتاه جمله در معیار bleu می‌باشد (با افزایش مقدار آلفا) تا ماشین ممانعت بیشتری در برابر کوتاه شدن جملات خروجی نشان دهد.

۲. اصطلاحات مربوط به ساعت

به دو نمونه زیر توجه کنید :

نمونه شماره 218 :

we will be back in Hamburg at five past ten .

ترجمه های مرجع :

0.ما ده و پنج دقیقه در هامبورگ خواهیم بود .

1.ما ده و پنج دقیقه در هامبورگ خواهیم بود .

2.ما دهوپنج دقیقه به هامبورگ بازخواهیمگشت .

3.ما ساعت ده و پنج دقیقه در هامبورگ خواهیم بود .

ترجمه ماشین :

ما در هامبورگ حاضر حدود ده و نیم .

==========

نمونه شماره 219 :

yes , we will be back at Hamburg at five past ten . would you like to go to the Wienerwald in Hanover in the evening ?

ترجمه های مرجع :

0.بله ، ما ده و پنج دقیقه در هامبورگ خواهیم بود . آیا دوست داری عصر به وینروالد در هانوور بروی ؟

1.بله ، ما ده و پنج دقیقه به هامبورگ برخواهیم گشت . مایلید بعدازظهر به وینروالد در هانوور برویم ؟

2.بله , ما دهوپنج دقیقه به هامبورگ بازخواهیمگشت . دوست داری که بعدازظهر به وینروالد در هانوور بروی ؟

3.بله ، ما ساعت دهوپنج دقیقه در هامبورگ خواهیم بود . دوست داری بعدازظهر در وینرواد هانوور باشی ؟

ترجمه ماشین :

بله ، ما ساعت پنج و نیم در هامبورگ خواهیم بود . آیا شما دوست دارید تا با هم در Hanover ؟

==========

مشاهده می‌شود که بارز ترین ایراد دو ترجمه‌ی بالا عدم توانایی فهم معنای «۵ دقیقه بعد از ده» توسط ماشین است و آن‌را با اعداد اعشاری اشتباه گرفته و «نیم» ترجمه می‌کند. شاید راهکار حل این مشکل وارد کردن دادگان بیشتری مربوط به ساعت باشد.

۲. تکرار لغات

به نمونه زیر توجه کنید :

نمونه شماره 222 :

I would like to take the train on the thirtieth .

ترجمه های مرجع :

0.من دوست دارم قطار روز سیام را بگیرم .

1.من مایلم سیام قطار بگیرم .

2.من دوست دارم که یک قطار برای سیام بگیرم .

3.دوست دارم در سیزدهم قطار بگیرم .

ترجمه ماشین :

من مایل هستم از سیام تا یک سیزدهم .

==========

مشکلی که در ترجمه‌ی بالا مشاهده می‌شود، حضور همزمان لغات «سی‌ام» و «سیزدهم» در ترجمه است درحالی که فقط یک تاریخ در متن اصلی حضور دارد. دلیل این پیش‌آمد این است که لغات مقصد فارغ از یک‌دیگر با لغات مبدا مقایسه می‌شوند و تکرار آن‌ها به دقت کمک می‌کند. شاید جلوگیری از Assign شدن دو لغت مجزا در خروجی به یک لغت از ورودی بتواند این مشکل را برطرف کند.

- replace_unk (ت

نقش این پارامتر آن است که اگر به True مقداردهی شود، پس از اتمام ترجمه، Tokenهایی که در ترجمه به UKN تبدیل شده‌اند با کلمه‌ای در زبان مبدا که بیشترین توجه را به آن دارد جایگزین می‌شود(در جملات فارسی از کلمات انگلیسی استفاده می‌شود). درصورتی که این پارامتر را قرار ندهیم، در ترجمه حاصل عبارات UKN ظاهر خواهند شد.

- ث) پنج پارامتر تأثیر گذار

۱. layers

این پارامتر تعداد لایه‌های شبکه را تعیین می‌کند که بدیهتا افزایش آن باعث افزایش دقت و همچنین افزایش زمان و دشواری مرحله‌ی آموزش می‌شود.

۲. heads

تعداد headهای لایه‌ی Attention را مشخص می‌کند و افزایش آن باعث افزایش قدرت این لایه می‌شود و شبکه قادر خواهد بود ارتباطات بیشتری را در بین کلمات پیدا کند.

۳. learning_rate

مانند هر شبکه‌ی دیگری، گام آموزشی کوتاه‌تر باعث ایجاد توانایی وارد شدن هرچه بیشتر شبکه به جزئیات می‌شود و همچنین سرعت آموزش را پایین می‌آورد.

۴. tgt_word_vec_size

با توجه به اینکه زبان مقصد ترجمه‌ی ما زبان فارسی است و می‌دانیم که لغات فارسی دارای جزئیات زیادی هستند، پس اندازه‌ی بردار بزرگ‌تر به ادراک بهتر لغات این زبان کمک می‌کند. از همین رو این پارامتر که اندازه‌ی بردار لغات زبان مقصد را مشخص می‌کند، می‌تواند تاثیرگذار باشد.

۵. encoder_type-decoder_type

این پارامترها، همانطور که نام آنها نشان می‌دهد، نوع شبکه‌های رمز نگار و رمز شکن را مشخص می‌کند که بنا بر زبان مبدا و مقصد می‌تواند بسیار تاثیرگذار باشد.

- ج) بررسی دو پارامتر

۱. tgt_word_vec_size - منبع : NLP_CA6_Part1_tgt_word_vec_size1000.ipynb

در این آزمایش، شبکه‌ی قبلی را با یک تغییر جزئی، دوباره مورد بررسی قرار دادیم. تغییر جزئی، تغییر اندازه بردارهای زبان مقصد (زبان فارسی) از 500 به 1000 است و دلیل این کار، همانطور که در بخش قبل توضیح داده شد، پیچیدگی لغات زبان فارسی است. پس از آموزش شبکه با شرایط یکسان با تجربه‌ی قبلی، معیار bleu بر روی دادگان تست به شکل زیر بدست آمد :

```
BLEU = 15.64, 56.7/24.6/11.5/4.7 (BP=0.942, ratio=0.943, hyp_len=2539, ref_len=2692)
BLEU = 17.60, 57.9/26.3/13.8/6.0 (BP=0.934, ratio=0.936, hyp_len=2539, ref_len=2713)
BLEU = 13.36, 51.0/19.7/9.4/4.3 (BP=0.939, ratio=0.940, hyp_len=2539, ref_len=2700)
BLEU = 14.83, 50.9/21.2/10.3/4.6 (BP=0.986, ratio=0.986, hyp_len=2539, ref_len=2575)
```

همانطور که انتظار می‌رفت، این تغییر باعث پیشرفت مدل شده و معیار افزایش پیدا کرده است. البته با توجه به اینکه افزایش اندازه‌ی بردارها به معنای افزایش تعداد پارامترهاست، انتظار افزایش زمان آموزش می‌رفت و این اتفاق افتاد و زمان آموزش در حدود دو برابر مقدار اولیه شد.

۲. layers - منبع : NLP_CA6_Part1_layers7.ipynb

در این آزمایش، شبکه‌ی قبلی را با یک تغییر جزئی، دوباره مورد بررسی قرار دادیم. تغییر جزئی، تغییر تعداد لایه‌هیا شبکه از 2 به 7 است و دلیل این کار، همانطور که در بخش قبل توضیح داده شد، افزایش توانایی یادگیری پیچیدگی‌های جملات است. بر خلاف انتظار، این تغییر باعث پسرفت شبکه شده و معیار bleu را برای آن کاهش داد. این مشاهده را می‌توان اینگونه توجیه کرد که تعداد لایه‌های بیشتر شبکه، تعداد پارامترهای نیازمند آموزش بیشتری نیز به همراه دارد و این موضوع زمان و دادگان بیشتری در مرحله‌ی آموزش نیاز دارد. همانطور که در فایل ضمیمه شده مشخص است، مشاهده می‌شود که شبکه به همگرایی می‌رسد، پس افزایش تعداد گام‌های آموزش کمکی نمی‌کند و کاستی از جهت کمبود دادگان است. همچنین طبق انتظار قبلی، با توجه به پارامترهای بیشتر، زمان آموزش شبکه نیز بیشتر شد.

زیر بخش دوم) با استفاده از bpe

- الف) نقش bleu - منبع :

نقش bpe در ترجمه، برخورد مناسبتر با لغات جدید مبنی بر زیر لغات آشنا است. در این حالت، برای شناخت لغات از خود لغات استفاده نمی‌شود و درک جمله در سطح زیر کلمات انجام می‌شود. در این صورت اسکیل مقدار پارامتر اندازه بردار کلمات متفاوت می‌شود چرا که تعداد زیر کلمات از تعداد کلمات بسیار محدودتر بوده و همچنین، هر زیر کلمه مفهوم کمتری از یک لغت در بر می‌گیرد، بنابراین، شبکه به اندازه‌ی بردار کوچکتری نیاز دارد. ما در این آزمایش اندازه بردار را تغییر ندادیم که به معنای توجه بیشتر به شناخت بهتر اجزا است.

- ب) معیار bleu بر روی دادگان ارزیابی - منبع : NLP_CA6_Part1_bpe.ipynb

پس از ۵۰۰ iteration آموزش شبکه، معیار bleu برای دادگان تست و چهار مرجع ترجمه به شکل زیر بدست آمد (به ترتیب مراجع) :

```
BLEU = 16.89, 56.7/25.2/12.4/5.7 (BP=0.948, ratio=0.949, hyp_len=2556, ref_len=2692)
BLEU = 19.03, 59.0/28.1/14.5/7.0 (BP=0.940, ratio=0.942, hyp_len=2556, ref_len=2713)
BLEU = 13.83, 51.6/20.9/9.3/4.6 (BP=0.945, ratio=0.947, hyp_len=2556, ref_len=2700)
BLEU = 15.41, 51.1/21.5/10.5/5.0 (BP=0.993, ratio=0.993, hyp_len=2556, ref_len=2575)
```

همانطور که مشاهده می‌شود، طبق انتظار، به دقت بالاتری رسیدیم. البته زمان آموزش حدودا دو برابر بود.

- ج) مقایسه دستی - منبع : NLP_CA6_Part1_bpe.ipynb

با کاوش در پاسخ‌های مدل به دادگان تست (بعد از iteration آخر)، یک پیشرفت و یک پسرفت در ترجمه‌ها یافتیم :

۱. بهبود

نمونه شماره 178 :

```
then we arrive at the airport in Hamburg again . right ?
```

ترجمه های مرجع :

0.پس دوباره به فرودگاه هامبورگ میرسیم . درست است .
1.پس ما دوباره به فرودگاه در هامبورگ میرسیم . درست است ؟
2.بعد ما دوباره به فرودگاه در هامبورگ میرسیم . درست ؟
3.پس ما دوباره به فرودگاه هامبورگ میرسیم . بسیارخب .

ترجمه ماشین بدون bpe :
پس آیا در فرودگاه ما خواهد رسید . درست است ؟

ترجمه ماشین با bpe :
پس ما باید دوباره در هامبورگ هستم . درست است ؟

تغییر امتیاز :
PRED SCORE: -2.7443 -> PRED SCORE: -1.9622

==========

مشاهده می‌شود که در این نمونه، مترجم ماشینی توانسته با استفاده از bpe کلمه‌ی ناآشنای هامبورگ که نام یک شهر است را ترجمه کند در حالی که قبلا، بدون استفاده از bpe قادر به این تشخیص نبوده است.

۲. پسرفت

نمونه شماره 22 :
What did you say ?

ترجمه های مرجع :
0.برای عصر برنامه چیست ؟
1.برای بعدازظهر برنامه چیست ؟
2.برای عصر چه برنامه‌های گذاشته شده است ؟
3.برای بعدازظهر چه نقشه‌ای داریم ؟

ترجمه ماشین بدون bpe :
برای عصر چه چیزی برنامه‌ریزی شده است ؟

ترجمه ماشین با bpe :
برای عصر برنامه‌ریزی میکنیم ؟

تغییر امتیاز :
PRED SCORE: -1.1049 -> PRED SCORE: -1.4283

==========

مشاهده می‌شود که مدل با bpe قدرت کمتری در تشخیص شناسه‌ها و زمان افعال دارد. این موضوع می‌توان معلول این موضوع باشد که وقتی اجزای جمله در سطح زیر کلمه بررسی می‌شوند، فاصله‌ی بین اجزای به هم مرتبط افزایش پیدا می‌کند و حفظ حافظه‌ی دنباله‌ای و پیدا کردن درست Attention دشوارتر می‌شود. بنابراین تشخیص موضوعاتی مثل شناسه و زمان فعلا که به Attention وابسته‌اند دشوار می‌شود.

بخش دوم) نویسه‌گردانی فارسی به انگلیسی

دادگان

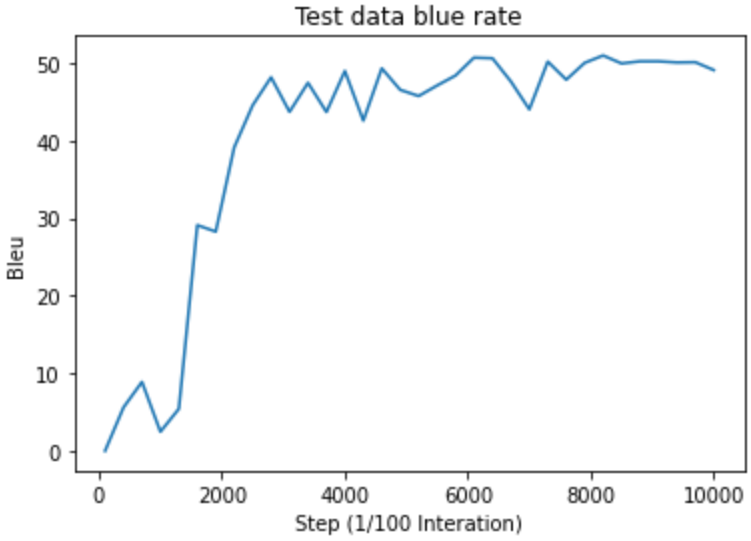
دادگان این بخش، سه دسته دادگان train, test, dev می‌باشد که هر دسته شامل دو فایل متنی موازی از جملات فارسی و نویسه گردانی شده به انگلیسی معادل است.

- الف) معیار bleu - منبع : NLP_CA6_Part2_default.ipynb

مقدار معیار bleu بر روی دادگان ارزیابی، بعد از ۱۰۰۰۰ گام آموزشی (iteration 5000) به شکل زیر حاصل شد :

```
BLEU = 10.19, 73.8/41.5/16.4/5.9 (BP=0.437, ratio=0.547, hyp_len=58074, ref_len=106171)
```

نمودار مقادیر این معیار بر روی دادگان توسعه، بر حسب گام آموزشی به شکل زیر به دست آمد (دقت این نمودار ۳۰۰ گام معدل iteration ۳ است) :



- ب) تحلیل معیار bleu

این معیار برای این مسئله مناسب نیست زیرا برحسب تطابق کامل کلمات عمل می‌کند، درحالی که واقعیت آن است که یک لغت فارسی، یک نویسه معادل انگلیسی مشخص و قطعی ندارد. از این رو داشتن یک نویسه مشابه آنچه انتظار می‌رود نیز مطلوب است اما bleu این موضوع را درنظر نمی‌گیرد.

- ج) معیارهای TER و WER

برای این مسئله، شاید معیارهای TER و WER مناسب‌تر باشند، چرا که این دو معیار بر حسب میزان تغییرات لازم (در سطح حروف) برای تبدیل کلمات به آنچه انتظار می‌رود محاسبه می‌شوند و باعث می‌شوند نویسه‌های مشابه نیز تا حدی مطلوب

باشند. در واقع هر معیاری که در مسئله‌ی اصلاح املایی کمک می‌کند می‌تواند در این مسئله نیز مفید باشد. متاسفانه به دلیل محدودیت RAM موفق به پیاده سازی محاسبه‌ی این معیارها نشدیم.

- ج) روش bpe - منبع : NLP_CA6_Part2_bpe.ipynb

انتظار می‌رفت که روش bpe در این مسئله کمک شایانی نکند چراکه به خودی خود، ما این مسئله را در سطح حروف بررسی می‌کنیم و زیربخشی برای این جزء وجود ندارد که در آن سطح بررسی شود. از همین رو انتظار می‌رفت که اعمال این روش هیچ کمکی نکند. با امتحان کردن این روش در فایلی که ارجاع داده شده است، نتیجه مورد انتظار را مشاهده کردیم و تغییری مشاهده نشد. احتمالا اگر bpe را در سطح کلمه اعمال کرده و با آن مسئله را حل کنیم کمک بیشتری بکند، چرا که چند حرفی‌های پشت سر هم معادل‌های نویسه‌ای مشهوری دارند و bpe در سطح کلمه ممکن است بتواند این مسئله را اعمال کند.

فایل‌های جانبی

به همراه این گزارش، یک پوشه به نام Codes ارائه می‌شود که حاوی فایل‌های زیر است :

- NLP_CA6_Part1_bpe.ipynb
- NLP_CA6_Part1_default.ipynb
- NLP_CA6_Part1_layers7.ipynb
- NLP_CA6_Part1_tgt_word_vec_size1000.ipynb
- NLP_CA6_Part2_bpe.ipynb
- NLP_CA6_Part2_default.ipynb

-