به نام پروردگار

پردازش زبان طبیعی تمرین کامپیوتری ۶

موعد تحویل : ۷ خرداد ۹۹

آزادی (ft.azadi@gmail.com)

در این تمرین قصد داریم با مراحل آموزش یک مترجم ماشینی مبتنی بر شبکه عصبی و پارامترهای موثر بر آن آشنا شویم. برای این کار در این تمرین از ابزار openNMT-py استفاده می کنیم، که ابزاری متن باز برای ترجمه ماشینی مبتنی بر شبکه عصبی است و مدل های sequence to sequence مختلف در آن پیاده‌سازی شده‌اند. این ابزار علاوه بر ترجمه ماشینی برای تسک های دیگری نظیر image to text، speech to text، خلاصه‌سازی و ... هم می تواند کاربرد داشته باشد.

برای اطلاع از نحوه نصب و استفاده از این ابزار و همچنین آشنایی با پارامترهای آن می توانید از دو لینک زیر استفاده کنید.

https://opennmt.net/OpenNMT-py/main.html

https://colab.research.google.com/drive/۱Nkd۹UFlDX۴NhX_gVQwDS-۷۷s۲jV۷zTqE#scrollTo=ZdTjS۰bTSVLk

دو تسک ترجمه انگلیسی به فارسی و نویسه‌گردانی[1] فارسی به انگلیسی در این تمرین مدنظر قرار گرفته‌اند، که باید این دو سیستم را آموزش داده و به سوالات مطرح شده پاسخ دهید.

آموزش سیستم ترجمه انگلیسی به فارسی

برای این تسک پیکره کوچکی انتخاب شده که همراه این تمرین ارسال شده است. در این بخش می خواهیم با استفاده از این پیکره یک سیستم ترجمه‌ی انگلیسی به فارسی آموزش دهیم.

---

[1] Transliteration

برای آموزش و تست یک سیستم ترجمه با OpenNMT-py باید مراحل زیر را انجام دهید. هر کدام از این مراحل پارامترهای خاص خود را دارد که می‌توان آن ها را تنظیم کرد. می توانید با رجوع به لینک بالا با این پارامترها آشنا شوید.

- پیش پردازش داده ها:

```
- python OpenNMT-py/preprocess.py -train_src "src-file"
     -train_tgt "tgt-file" -valid_src "src-dev-file"
     -valid_tgt "tgt-dev-file" -save_data "data-out-file"
```

- آموزش مدل:

```
- python OpenNMT-py/train.py -data "data-out-file"
     -save_model "model-address/model" -world_size 1 -gpu_rank 0
     -train_steps 50000
```

- ترجمه فایل تست:

```
- python OpenNMT-py/translate.py
     -model "model-address/model_step_50000.pt"
     -src "test-file" -output "output-address" -replace_unk -verbose
```

- ارزیابی خروجی:

```
- perl  OpenNMT-py/tools/multi-bleu.perl "reference-address" < "output-
  address"
```

۱- یک سیستم ترجمه مبتنی بر RNN با پارامترهای پیشفرض و بدون استفاده از bpe آموزش دهید.

الف) معیار BLEU برای سیستم ترجمه‌ی آموزش داده شده بعد از ۵۰۰۰۰ iteration را روی خروجی پیکره تست با ۴ مرجع محاسبه کنید.

ب) نمودار مقدار BLEU بر حسب تعداد iteration روی مجموعه dev را رسم نموده و بررسی کنید مقدار BLEU روی مجموعه dev تقریبا بعد از چند iteration همگرا می‌شود. (برای این کار می توانید از پارامتر save_checkpoint_steps استفاده کنید و مدل های میانی را در حین آموزش ذخیره کنید.)

پ) دو نمونه خطا در ترجمه های تولید شده توسط سیستم ترجمه پیدا کنید. برای هرکدام جمله‌ی مبدا، جمله‌ی مرجع و جمله‌ی تولیدی توسط سیستم ترجمه را در گزارش خود بیاورید. به نظر شما هر یک از این خطاها به چه دلیل ممکن است ایجاد شده باشد؟ برای هر یک راه حلی که به ذهنتان میرسد و ممکن است با ایجاد آن تغییر در سیستم ترجمه این مشکل حل شود را بیان کنید.

ت) پارامتر replace_unk در زمان ترجمه چه نقشی دارد؟ اگر آن را در زمان ترجمه قرار ندهیم در خروجی چه تغییری ایجاد می شود؟

ث) پارامترهای مربوط به آموزش مدل ترجمه را از این لینک بررسی کنید و ۵ پارامتری که به نظر شما تاثیر زیادی روی دقت نهایی سیستم ترجمه و یا سرعت همگرایی آن میتوانند داشته باشند را ذکر کنید. توضیح دهید این پارامترها هر کدام چه هستند و چه نقشی دارند.

ج) از پارامترهای بند قبل ۲ پارامتری که حس می کنید بیشترین تاثیر را می توانند داشته باشند انتخاب کرده و با تغییر آن ها در زمان آموزش شبکه تاثیر آن ها در دقت نهایی و مدت زمان آموزش شبکه را مورد بررسی قرار دهید.

۲- یک سیستم ترجمه مبتنی بر RNN با پارامترهای پیشفرض و با استفاده از bpe آموزش دهید.

برای آموزش با استفاده از مدل bpe باید ابتدا پیش از مرحله‌ی preprocess پیکره، کدهای bpe را از روی پیکره آموزشی یادگرفته و سپس روی همه‌ی پیکره ها اعمال کنید. پیش از ارزیابی نیز باید کلماتی که با bpe شکسته شده‌اند را مجددا به هم متصل کنید و BPE Detokenization انجام دهید. برای آشنایی با نحوه انجام این کار به دو لینکی که در ابتدای تمرین قرار داده شده است مراجعه کنید. پس از آموزش مترجم با استفاده از bpe به سوالات زیر پاسخ دهید.

الف) توضیح دهید استفاده از bpe چه نقشی در آموزش سیستم ترجمه می‌تواند داشته باشد؟ چه پارامتری در آموزش شبکه در این جا با آزمایش قبلی تفاوت پیدا می کند؟

ب) معیار دقت بلو برای سیستم ترجمه‌ی جدید را محاسبه کرده و با حالت قبلی مقایسه کنید.

ج) یک نمونه از ترجمه هایی که در سیستم جدید نسبت به قبلی بهتر شده و یک نمونه ترجمه ای که بدتر شده پیدا کنید و به همراه جمله مبدا و جمله‌ی مرجع آنها ذکر کنید. توضیح دهید که به نظر شما چرا ترجمه با استفاده از bpe در نمونه ای که یافته اید بهتر یا بدتر شده است.

آموزش سیستم نویسه‌گردانی فارسی به انگلیسی

در این بخش می خواهیم با استفاده از پیکره های داده شده برای نویسه‌گردانی یک سیستم نویسه‌گردانی فارسی به انگلیسی آموزش دهیم. پیکره های موجود در پوشه‌ی Transliteration شامل پیکره های train، dev و test است، که شامل جملات فارسی و حالت نویسه‌گردانی شده آن به زبان انگلیسی (فینگلیش) است.

سیستم نویسه‌گردانی را به‌صورت یک سیستم ترجمه در سطح کاراکتر مدل می‌کنیم. بنابراین لازم است که ابتدا یک پیش‌پردازش روی داده ها انجام داده و کاراکترهای موجود در جمله را با یک فاصله از هم جدا کنید. توجه کنید که کاراکترهای فاصله‌ی موجود در بین کلمات نیز باید به عنوان یک کاراکتر مستقل در نظر گرفته شوند و برای اینکه از فاصله های اضافه شده بین حروف متمایز شوند، آن ها را نیز باید با سمبلی مانند <b> جایگزین کنید. برای نمونه جمله‌ی "من رفتم" پس از پیش پردازش به جمله‌ی "م ن <b> ر ف ت م" تبدیل می شود.

پس از آماده‌سازی داده‌ها، مشابه سیستم ترجمه بدون استفاده از bpe در بخش قبل، سیستم نویسه‌گردانی فارسی به انگلیسی را نیز آموزش دهید و سپس به سوالات زیر پاسخ دهید.

الف) معیار بلو برای سیستم آموزش داده شده روی خروجی پیکره تست را به دست آورید.

ب) با ذکر دلیل توضیح دهید که به نظر شما آیا معیار بلو برای ارزیابی تسک نویسه گردانی مناسب است؟

ج) یک معیار برای ارزیابی نویسه گردانی پیشنهاد دهید که به نظرتان بهتر از معیار بلو باشد، سپس با معیار ارائه شده‌ی خودتان خروجی سیستم آموزش داده شده روی پیکره‌ی تست را ارزیابی کنید. چرا به نظرتان این معیار بهتر از بلو بوده است؟

د) آیا استفاده از روش bpe در این تسک می‌تواند مفید باشد؟ چرا؟

لطفا علاوه بر فایل گزارش، فایل اسکریپت دستورات اجرا شده و یا اگر در colab اجرا کرده‌اید فایل notebook آن به همراه خروجی سیستم های آموزش داده شده برای فایل تست را نیز ارسال کنید.

لطفا به قواعد حل تمرین که در CECM قرار داده شده است توجه کنید.