

به نام خدا



دانشگاه تهران

دانشکده فنی

دانشکده مهندسی برق و کامپیوتر



درس پردازش زبان طبیعی

پاسخ تمرین ۵

سید محمد مهدی رضوی

۸۱۰۱۰۲۱۵۵

خرداد ماه ۱۴۰۳

فهرست

فهرست تصاویر.....	۲
پاسخ سوال اول.....	۳
پاسخ بخش اول.....	۳
پاسخ بخش دوم.....	۷
پاسخ بخش سوم.....	۸
پاسخ بخش چهارم.....	۸

فهرست تصاویر

شکل ۱ : تعداد کل خطوط دینای فارسی و انگلیسی و ۳ سطر اول آن‌ها.....	۳
شکل ۲ : کامند برای پیش‌پردازش مترجم ماشینی انگلیسی به فارسی.....	۴
شکل ۳ : نمودار هیستوگرام کلمات انگلیسی.....	۴
شکل ۴ : نمودار هیستوگرام کلمات فارسی.....	۵
شکل ۵ : هایپرپارامترهای اولیه برای آموزش مدل های ترنسفورمر.....	۵
شکل ۶ : اجرای مدل ترنسفورمر بر روی کامپیوتر شخصی.....	۶
شکل ۷ : نتیجه اجرای یک اپیک بر روی کامپیوتر شخصی.....	۷
شکل ۸ : مقدارامتیاز بلو برای ال اس تی ام.....	۷
شکل ۹ : ماکزیمم امتیاز بلو به دست آمده.....	۹

پاسخ سوال اول

پاسخ بخش اول

یک میلیون و بیست و یک هزار سطر داده فارسی و انگلیسی موازی در دیتاست موجود است که پس از فیلتر کردن ، پانصد هزار سطر داده موجود خواهد بود.

```
File Edit View Search Terminal Help
mahdi@MahdiRazavi: ~/Documents/NlpUt/NLP_CA5_810102155$ ./analysis.sh
before filtering
-----
1021597 ./data/en-fa.txt/MIZAN.en-fa.en
1021597 ./data/en-fa.txt/MIZAN.en-fa.fa
-----
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.
داستانی که از نظر شما میگذرد، ابتدا ضمن کنفرانس صلح پاریس از روی یادداشتهایی 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
و از روی گزارشاتی که برای رؤسای من در قاهره ارسال گردیده بودند نوشته شد
بعداً در پائیز سال 1919، این نوشته اولیه و بعضی از یادداشتهای، مفقود شدند
-----
after filtering
-----
548185 ./filtered data/train.fa
548185 ./filtered data/train.en
-----
داستانی که از نظر شما میگذرد، ابتدا ضمن کنفرانس صلح پاریس از روی یادداشتهایی 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
و از روی گزارشاتی که برای رؤسای من در قاهره ارسال گردیده بودند نوشته شد
بعداً در پائیز سال 1919، این نوشته اولیه و بعضی از یادداشتهای، مفقود شدند
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.
mahdi@MahdiRazavi:~/Documents/NlpUt/NLP_CA5_810102155$
```

شکل ۱: تعداد کل خطوط دیتای فارسی و انگلیسی و ۳ سطر اول آنها

در تصویر زبان دستور مربوط به عملیات پیش پردازش داده برای واژگان به اندازه ده هزار کلمه را شاهد هستیم.

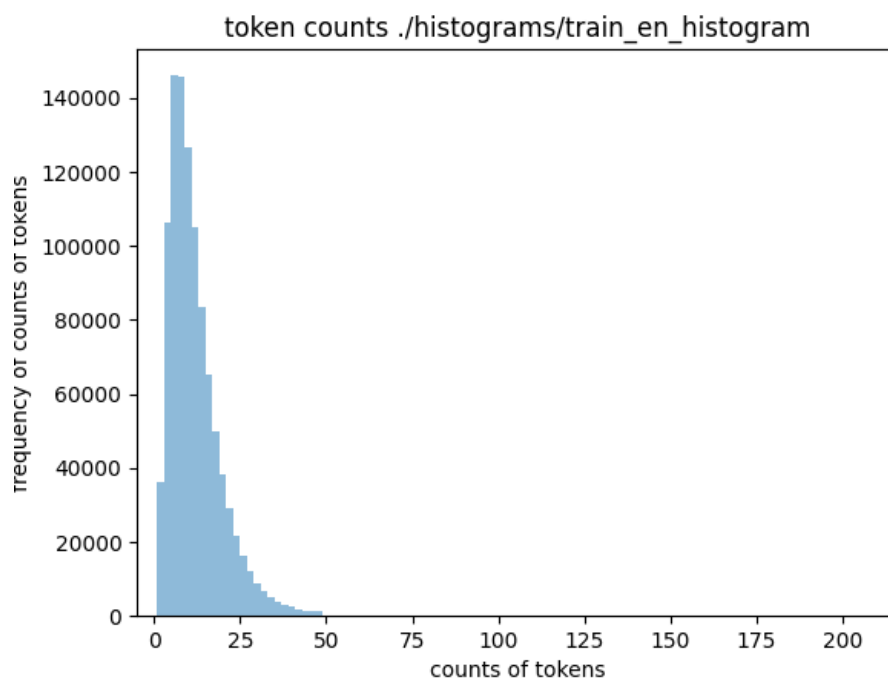
```

fairseq-preprocess --source-lang en --target-lang fa \
--trainpref ../final_data/train/train \
--validpref ../final_data/valid/valid \
--testpref ../final_data/test/test \
--destdir ./data_bin/ --nwordstgt 10000 --nwordssrc 10000 \
--workers 20

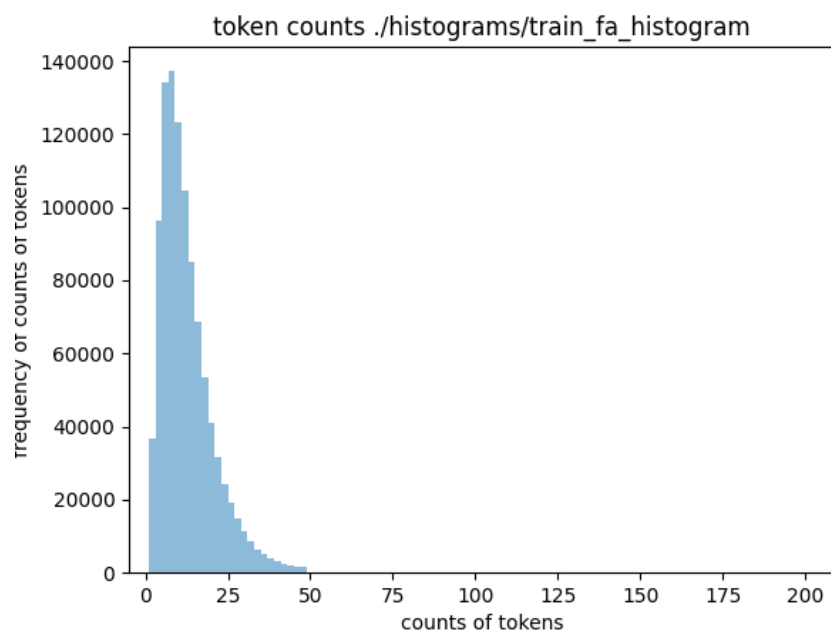
```

شکل ۲: کامند برای پیش پردازش مترجم ماشینی انگلیسی به فارسی

بیشترین نرخ فراوانی کلمات در سطرها را، برای سطرها با ۲۰ کلمه مشاهده خواهیم کرد.



شکل ۳: نمودار هیستوگرام کلمات انگلیسی



شکل ۴: نمودار هیستوگرام کلمات فارسی

Telemetry Consent	Welcome	Sunsetting Atom	encoderdecodertransfo...	encoderdecoderlstm.sh	encoderdecodertransformer.sh
1	export CUDA_VISIBLE_DEVICES=0				1 fairseq-train ../preprocess/data_bin/ \
2					2 --cpu \
3	fairseq-train ../preprocess/data_bin/ \				3 --task translation \
4	--cpu \				4 --arch transformer \
5	--task translation \				5 --encoder-layers 6 \
6	--arch lstm \				6 --decoder-layers 6 \
7	--encoder-layers 6 \				7 --encoder-embed-dim 512 \
8	--decoder-layers 6 \				8 --decoder-embed-dim 512 \
9	--encoder-hidden-size 512 \				9 --encoder-ffn-embed-dim 2048 \
10	--decoder-hidden-size 512 \				10 --decoder-ffn-embed-dim 2048 \
11	--max-tokens 4000 \				11 --encoder-attention-heads 8 \
12	--max-sentences 32 \				12 --decoder-attention-heads 8 \
13	--criterion label_smoothed_cross_entropy \				13 --max-tokens 4000 \
14	--label-smoothing 0.1 \				14 --max-sentences 32 \
15	--dropout 0.2 \				15 --dropout 0.2 \
16	--optimizer adam \				16 --optimizer adam \
17	--lr 0.0005 \				17 --lr 0.0005 \
18	--lr-scheduler inverse_sqrt \				18 --lr-scheduler inverse_sqrt \
19	--warmup-updates 4000 \				19 --warmup-updates 4000 \
20	--weight-decay 0.0001 \				20 --weight-decay 0.0001 \
21	--clip-norm 0.1 \				21 --clip-norm 0.1 \
22	--save-dir ./EncoderDecoderLSTM/ \				22 --save-dir ./encoderdecodertransformer/
23					23

شکل ۵: هایپرپارامترهای اولیه برای آموزش مدل های ترنسفورمر

تلاش های بسیار زیادی در این تمرین انجام دادم تا بتوانم بر روی پی یو کامپیوتر شخصی خودم اجرا بگیرم ، اما متاسفانه درایورهای پردازش موازی در کامپیوتر شخصی بنده موجود نبود ، با این حال اجرای یک ایپاک از این مدل ترنسفورمر حدود ۵ ساعت از من زمان گرفت . در زیر تصاویر مربوط به این اجرا آورده شده است .

```

mahdi@MahdiRazavi: ~/Documents/NLP/CAS_810102155/EncoderDecoder_LSTM
File Edit View Search Terminal Help

(encoder): LSTMEncoder(
  (dropout_in_module): FairseqDropout()
  (dropout_out_module): FairseqDropout()
  (embed_tokens): Embedding(10000, 512, padding_idx=1)
  (lstm): LSTM(512, 512, num_layers=6, dropout=0.2)
)
(decoder): LSTMDecoder(
  (dropout_in_module): FairseqDropout()
  (dropout_out_module): FairseqDropout()
  (embed_tokens): Embedding(10000, 512, padding_idx=1)
  (layers): ModuleList(
    (0): LSTMCell(1024, 512)
    (1-5): 5 x LSTMCell(512, 512)
  )
  (attention): AttentionLayer(
    (input_proj): Linear(in_features=512, out_features=512, bias=False)
    (output_proj): Linear(in_features=1024, out_features=512, bias=False)
  )
  (fc_out): Linear(in_features=512, out_features=10000, bias=True)
)
2024-06-07 10:51:56 INFO | fairseq_cli.train | task: TranslationTask
2024-06-07 10:51:56 INFO | fairseq_cli.train | model: LSTMModel
2024-06-07 10:51:56 INFO | fairseq_cli.train | criterion: LabelSmoothedCrossEntropyCriterion
2024-06-07 10:51:56 INFO | fairseq_cli.train | num. shared model params: 42,419,984 (num. trained: 0)
2024-06-07 10:51:56 INFO | fairseq_cli.train | num. expert model params: 0 (num. trained: 0)
2024-06-07 10:51:56 INFO | fairseq.data.data_utils | loaded 5,000 examples from: ../preprocess/data/bin/valid.en-fa.en
2024-06-07 10:51:56 INFO | fairseq.data.data_utils | loaded 5,000 examples from: ../preprocess/data/bin/valid.en-fa.fa
2024-06-07 10:51:56 INFO | fairseq.tasks.translation | ../preprocess/data/bin/valid.en-fa 5000 examples
2024-06-07 10:51:56 INFO | fairseq.trainer | detected shared parameter: decoder.attention.input_proj.bias <- decoder.attention.output_proj.bias
2024-06-07 10:51:56 INFO | fairseq_cli.train | training on 1 devices (GPUs/TPUs)
2024-06-07 10:51:56 INFO | fairseq_cli.train | max tokens per device = 4000 and max sentences per device = 32
2024-06-07 10:51:56 INFO | fairseq.trainer | Preparing to load checkpoint ../EncoderDecoderLSTM/checkpoint_last.pt
2024-06-07 10:51:56 INFO | fairseq.trainer | No existing checkpoint found ../EncoderDecoderLSTM/checkpoint_last.pt
2024-06-07 10:51:56 INFO | fairseq.trainer | loading train data for epoch 1
2024-06-07 10:51:56 INFO | fairseq.data.data_utils | loaded 500,000 examples from: ../preprocess/data/bin/train.en-fa.en
2024-06-07 10:51:56 INFO | fairseq.data.data_utils | loaded 500,000 examples from: ../preprocess/data/bin/train.en-fa.fa
2024-06-07 10:51:56 INFO | fairseq.tasks.translation | ../preprocess/data/bin/train.en-fa 500000 examples
2024-06-07 10:51:57 INFO | fairseq.data.iterators | grouped total_num_itr = 15627
epoch 001: 0% | 0/15627 [00:00<?, ?it/s]2024-06-07 10:51:57 | INFO | fairseq
.trainer | begin training epoch 1
2024-06-07 10:51:57 | INFO | fairseq_cli.train | Start iterating over samples
epoch 001: 9% | 1429/15627 [31:27<5:25:46, 1.38s/it, loss=9.759, nll loss=9.181, ppl=580.63, wps=298.6, ups=0.67, wpb=]

```

شکل ۶: اجرای مدل ترنسفورمر بر روی کامپیوتر شخصی

[illegible]

شکل ۹ : نتایج حاصل از مدل انکودر دیکودر با ترنسفورمر

همانطور که در تصویر بالا نیز مشاهده می‌شود، در این مدل استفاده از رویکرد حریصانه موجب گردیده است که مدل دایما کلمات تکراری که معادل با کلمه اول تولید شده است را تولید کند، چون جستجوی بیم، هر دفعه کلمه با بیشترین احتمال ممکنه را به عنوان کلمه بعدی به عنوان خروجی منتشر خواهد کرد.

مقدار امتیاز بلو این مدل با گذشت ایام‌های بیشتر تغییر نمی‌کند و در مقدار صفر ثابت مانده‌است.

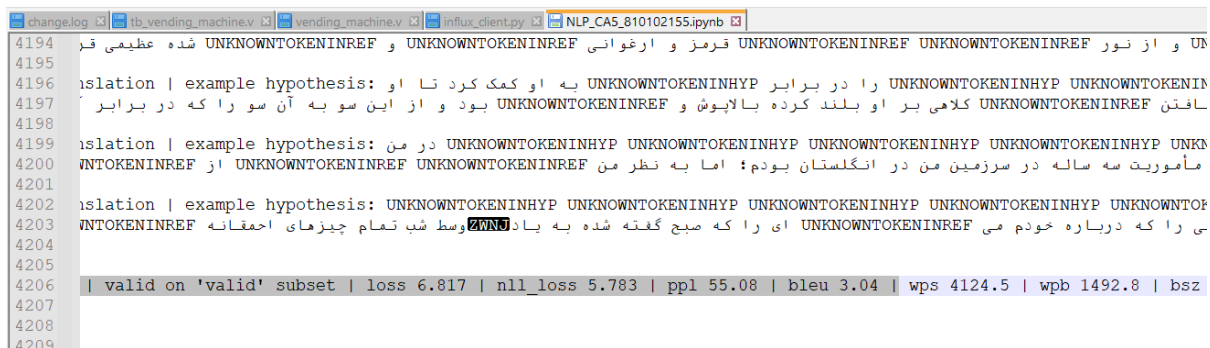
COMET کیفیت ترجمه را در یک مفهوم کلی اندازه گیری می کند. COMET لزوماً پس از سفارشی سازی افزایش نمی یابد (سیستم سفارشی شده ممکن است ترجمه هایی با کیفیت مشابه تولید کند، تفاوت این است که آیا ترجمه ها با سبک، لحن صدا، اصطلاحات و غیره مشتری مطابقت دارند یا خیر). با این حال، کاهش قابل توجه COMET ممکن است نشان دهنده یک مشکل در سیستم سفارشی شده باشد.

COMET از یک مدل شبکه عصبی آموزش داده شده بر روی مجموعه داده های بزرگ قضاوت های انسانی استفاده می کند. این ترجمه ها را با در نظر گرفتن جنبه های مختلف کیفیت ترجمه، از جمله روان بودن، کفایت، و حفظ معنا ارزیابی می کند.

BLEU، یکی از قدیمی ترین و پرکاربردترین معیارها، کیفیت متن ترجمه شده با ماشین را با مقایسه آن با یک یا چند ترجمه مرجع با کیفیت بالا ارزیابی می‌کند. BLEU مطابقت عبارات بین متن تولید شده توسط ماشین و متون مرجع را با تمرکز بر دقت تطابق کلمات می‌سنجد.

BLEU دقت n گرم را برای طول های مختلف n گرم (معمولاً ۱ تا ۴ کلمه) محاسبه می کند و سپس این امتیازها را با استفاده از میانگین هندسی ترکیب می کند. همچنین شامل جریمه اجاز برای پرداختن به موضوع ترجمه های بسیار کوتاه می شود.

BLEU به ویژه برای ارزیابی ترجمه‌هایی که تطابق دقیق عبارات و ترتیب کلمات مهم است، مؤثر است. با این حال، اتکای آن به تطابق دقیق می‌تواند محدودیتی در دریافت کیفیت ترجمه‌های روان یا اصطلاحی باشد.



The screenshot shows a Jupyter Notebook interface with several tabs at the top: 'change.log', 'tb_vending_machine.v', 'vending_machine.v', 'influx_client.py', and 'NLP_CA5_810102155.ipynb'. The active tab is 'NLP_CA5_810102155.ipynb'. The notebook content displays a translation example with a hypothesis and a reference, followed by a table of performance metrics. The metrics table is as follows:

valid on 'valid' subset	loss	nll_loss	ppl	bleu	wps	wpb	bsz
6.817	5.783	55.08	3.04	4124.5	1492.8	bsz	

شکل ۹: ماکزیمم امتیاز بلو به دست آمده

پس از ۱۳ ایپاک یادگیری به امتیاز بلو ۳,۰۴ مدل انکودر دیکودر دست پیدا کرد. همچنین متن فرض ترجمه با متن رفرنس تا حد بهتری تطابق خواهند داشت.