NLP

CA05

MohamadJavad Kamyab
810100457
Hossein Seifi
810100386
University of Tehran

فهرست

پیش پردازش ۲
ابزار open-nmt ۴
مقادیر و کانفیگ ٤
خروجی و نتایج ٦
ابزار Marian-NMT ۸
مقادیر و کانفیگ ۸

پیش پردازش

پیش پردازش انجام شده برای این بخش شامل موارد زیر است:

۱. کاراکترهایی مانند "ك" را به "ک" تبدیل کردن (نرمال‌سازی کاراکترها)
۲. پیدا کردن و تصحیح غلط‌های املایی در جمله
۳. و در نهایت توکنایز کردن جمله

پیش‌پردازش‌های دیگر مانند حذف علائم نگارشی، ریشه‌یابی کلمات و... ممکن است روند یادگیری ماشین ترجمه را تحت تأثیر قرار بدهد زیرا پس از ریشه‌یابی تنها کلمه‌ای که ماشین آن را فرا گرفته‌است ریشه کلمات است که در یک جمله تنها ریشه کلمات استاده نمی‌شوند، در نتیجه مدل نیز بخوبی آموزش داده نمی‌شود.

در زبان فارسی نرمال کردن کاراکترها اهمیت زیادی دارد چرا که کلمات گاهی با کیبوردهای با زبان غیر فارسی نوشته می‌شوند در نتیجه یکپارچگی جملات را از بین می‌برد. پس بهتر است که این نرمال‌سازی انجام شود.

برای زبان انگلیسی نیز مورد صدق می‌کند که اگر "O" به‌صورت "Ò" نوشته شده باشد!

ابزاری که برای رفع اشکال از نظر املایی برای زبان فارسی استفاده شده است ابزار parsivar می‌باشد و همتای آن برای زبان انگلیسی ابزار autocorrect نام دارد.

برای توکن کردن bpe از ابزار از پیش تعریف شده‌ای که درون open-nmt repository بود استفاده شد. این توکنایزر را بر روی دیتا train فارسی و انگلیسی موجود آموزش دادیم و سپس آن را بر روی تمامی مجموعه‌داده به غیر از خروجی آخر یعنی test.fa اعمال کردیم.

دیتایی که برای آموزش این توکنایزر استفاده شده بسیار کم بود در نتیجه عملکرد بسیارخوبی نداشت و این موضوع که آیا مجاز به افزودن داده برای آموزش توکنایزر هستیم یا خیر در صورت مسئله قید نشده بود با همین روال ادامه دادیم.

نمونه‌ای از داده‌ی آموزشی قبل و بعد از پیش‌پردازش به‌صورت زیر می‌باشد:

قبل از پیش‌پردازش:

eng	per
maybe hes just gone into wahine , meggie suggested .	مگی درآمد که: شاید فقط رفته به واهاین.
i assembled this from a corpse , remolding its flesh and bones .	ها رو از یه جسد برداشتم و ظاهر پوست و این استخوانش: می‌گوید رو عوض کردم.
the horse slipped on the wet grass	اسب روی علف‌های خیس لیز می‌خورد.
but there is a pale shade of bribery which is sometimes called prosperity .	اما یک نوع رشوهی دیکری هم هست که گاهی اسمش را موفقیت مالی می‌گذارند.
all too soon , the hogwarts express was pulling in at platform nine and three quarters .	قطار هاگوارتز نیز زودتر از آنکه انتظارش می‌رفت شروع به کم کردن سرعتش کرد و در ایستگاه نه و سه چهارم متوقف شد.
for this purpose , scientific information of iranian origin in form of articles dealing with basic sciences , were extracted . iran 's positionhas been compared with the rest of the world	برای این منظور تولید اطلاعات علمی ایرانیان در قالب نمایه‌نامه استنادی علوم " مقالات و در زمینه علوم پایه از " استخراج‌شده و وضعیت ایران با جهان مقایسه شده‌است.
while he slept took a rib from his left side	در آن دم که آدم در خواب بود از پهلوی چپ او دنده‌ای برداشت
take off my shirt ?	جان گفت: پیرهنمو در بیارم?
though goriot 's eyes seemed to have shrunk in their sockets	اگرچه پلک زیرین چشم‌های گوری و قدری ورم کرده و برگشته بود

his face brought back all my fear and then some .	صورتش همه ترس‌های مرا و چیز دیگری را به من برگرداند.

بعد از پیش‌پردازش:

eng	per
maybe hes just gone into white , maggie suggested .	ملی در آمد که: شاید فقط رفته به وا ه@@ این.
i assemb@@ led this from a corpse , recording its flesh and bones .	ها را از به جسد بر@@ داشتم و ظاهر پوست و است این خوان@@ ش: می‌گوید و عوض کردم.
the horse slipped on the wet grass	است روی عل@@ ف‌های خیس نیز می‌خورد.
but there is a pale shade of bri@@ ber@@ y which is sometimes called prosp@@ er@@ ity .	اما یک نوع رشوه ی دیگری هم است که گاهی اسمش را موفقیت مالی می‌گذارند.
all too soon , the hogwarts express was pulling in at platform nine and three quarters .	قطار هاگوارتز نیز زودتر از آنکه انتظار می‌رفت شروع به کم کردن سر@@ ع@@ تش کرد و در ایستگاه به و سه چهارم متوقف شد.
for this purpose , scientific information of iranian origin in form of articles dealing with basic sciences , were extrac@@ ted . iran 's positions been compared with the rest of the world	برای این منظور تولید اطلاعات علمی ایر@@ انسان در قالب نمای@@ ه نامه است@@ " مقالات و در زمینه علوم پایه از نا@@ دی علوم " استخرا@@ ج@@ شده و وضعیت ایران با جهان مقایسه شده‌است.
while he slept took a ri@@ b from his left side	در آن در که عدم در خواب بود از پهلوی چپ او دن@@ ده‌ای برداشت
take off my shirt ?	جان گفت: پی@@ ره نم@@ و در بی@@ ارم?
though riot 's eyes seemed to have sh@@ run@@ k in their soc@@ kets	اگرچه پلک زی@@ رین چشم‌های گو ریو قدرت و رم کرده و بر گشته بود
his face brought back all my fear and then some .	صورت همه تر@@ س‌های مرا و نیز دیگری را به من بر گرداند.

ابزار open-nmt

در این بخش پیاده‌سازی و عملکرد open-nmt را مورد بررسی قرار می‌دهیم

مقادیر و کانفیگ

در ابتدا به ساخت دیکشنری که مورد نیاز مدل است از روی داده‌های آموزش اقدام می‌کنیم.

```
!onmt-build-vocab --size 50000 --save_vocab /content/data/en-
vocab.txt /content/data/train.en.bpe
```

```
!python OpenNMT-tf/third_party/learn_bpe.py -i data/train.fa -
o data/fa.code -s 10000
```

این مشخصات فایل data.ymlی است که در درایو موجود است و فایل آن برای کانفیگ کردن مدل استفاده می‌شود که به تشریح آن‌ها می‌پردازیم:

در بخش اول که data نام دارد آدرس فایل‌های مورد نیاز آن یعنی فایل‌های train و valid و همچنین vocab مدل را مشخص می‌کنیم.

```
data:
  eval_features_file: /content/data/valid.en.bpe
  eval_labels_file: /content/data/valid.fa.bpe
  source_words_vocabulary: /content/data/en-vocab.txt
  target_words_vocabulary: /content/data/fa-vocab.txt
  train_features_file: /content/data/train.en.bpe
  train_labels_file: /content/data/train.fa.bpe
```

در بخش eval که مربوط به ارزیابی مدل در حین آموزش است پارامترهای زیر مشخص شده است.

```
eval:
  batch_size: 30
```

مشخص می‌کند بعد از چند نمونه مقدار را در شبکه‌ی عصبی propagated کند

```
  eval_delay: 1800
```

بعد از ثانیه ارزیابی سیستم را انجام دهد

```
  export_format: saved_model
```

مدلی که در خروجی ذخیره می‌شود به‌صورتی باشد

```
  export_on_best: bleu
```

ارزیابی مدل بر روی کدام یک از ارزیاب ها اتفاق بیفتد

عملاً این بخش مقدار validی که به‌دست می‌آید را با این ارزیاب اندازه‌گیری می‌کن و ملاک پیشرفت مدل را بر روی این مقدار قرار می‌دهد.

```
  exporters: last
```

کدام مدل آموزشی ذخیره شود

```
  length_bucket_width: 5
```

برای مقایسه کردن با ترجمه در مراحل قبلی استفاده می‌شود که اگر تا پنج مرحله قبل موجود باشد بتواند مقایسه را انجام دهد.

```
  max_exports_to_keep: 15
```

چند مقدار خروجی را برای ارزیابی نگهدارد

```
  save_eval_predictions: true
```

آیا مقدار پیش‌بینی شده در هر مرحله را نیز نگهداری کند

```
  scorers: bleu, rouge, wer, ter, prf
```
ارزیاب‌هایی که در هر مرحله مدل را مورد ارزیابی قرار دهند
```
  steps: 1000
```
تعداد گام‌ها برای ارزیابی
```
infer:
  batch_size: 128
  bucket_width: null

model_dir: /content/drive/MyDrive/NLP/CA05/run/
```
مکانی که مدل در آن ذخیره شود
```
params:
  beam_width: 5
```
پهنای جستجو در هر مرحله
```
  clip_gradients: 5.0
  decay_params:
    decay_rate: 0.7
    decay_steps: 100000
```
تنظیمات مربوط به decay که نشان می‌دهد داده‌ها پس از چند گام فراموش شوند
```
  decay_type: exponential_decay
  learning_rate: 0.5
```
سرعت آموزش که به‌صورت پیش‌فرض روی ۱ بود ولی با ذهنیتی که در مورد این پارامتر داشتیم (برای مثال در تسک‌ها و مدل‌های قبلی این عدد بسیار کوچک‌تر بود در حد یک‌هزارم!) در نتیجه آن را به ۰٫۵ تغییر دادیم
```
  maximum_iterations: 50
  optimizer: GradientDescentOptimizer
  param_init: 0.1
  start_decay_steps: 500000
```
از کدام گام به بعد فرآیند decay را آغاز کند (در مدلی که ما آموزش دادیم از این پارامتر استفاده نشده چرا که هیچ‌گاه مدل به گام ۵۰۰،۰۰۰ نمی‌رسد!)
```
score:
  batch_size: 64
train:
  batch_size: 128
```
تعداد نمونه را برای آموزش دادن مشخص می‌کند (بعد از چند نمونه مقادیر و وزن‌ها را آپدیت کند)
```
  batch_type: examples
  bucket_width: 1
  maximum_features_length: 512
```
بیشترین طولی که برای هر نمونه به‌عنوان ویژگی در نظر می‌گیرد
```
  maximum_labels_length: 512
```
بیشترین طولی که برای ویژگی خروجی در نظر می‌گیرد
```
  sample_buffer_size: -1
  save_checkpoints_steps: 2500
```
بعد از چند گام مدل را ذخیره کند
```
  save_summary_steps: 50
  train_steps: 50000
```
تعداد گام‌ها آموزش

این مدل مجموعاً در ۵۰ هزار گام آموزش داده‌شده و مدت آموزش آن به ۵ ساعت ۲۰ دقیقه رسید همچنین یکی از پارامترهایی که به‌دلیل محدودیت زمانی و سخت‌افزاری کم در نظر گرفته شد همین گام آموزش است و دو متغیر دیگر یکی batch_size و دیگری طول فیچرها

در ورودی و خروجی است که اولی باعث می‌شود مدل دیرتر وزن‌ها خود را به‌روزرسانی کند و دومی باعث می‌شود نتواند بخوبی از متن ویژگی‌های مورد نظر خود را خارج کند که این امر باعث می‌شود بعضی از موارد مهم از قلم بیفتد.
ولی تأثیرگذارترین پارامتر همان تعداد گام است
البته یکی دیگر از پارامترها learning_rate بود که سرعت آموزش مدل را تعیین می‌کند اگر این عدد کمتر باشد فرآیند آموزش کندتر پیش می‌رود ولی از طرفی باعث می‌شود مدل خیلی جهش‌ها بزرگ خوب بد در فرآیند یادگیری نداشته باشد!

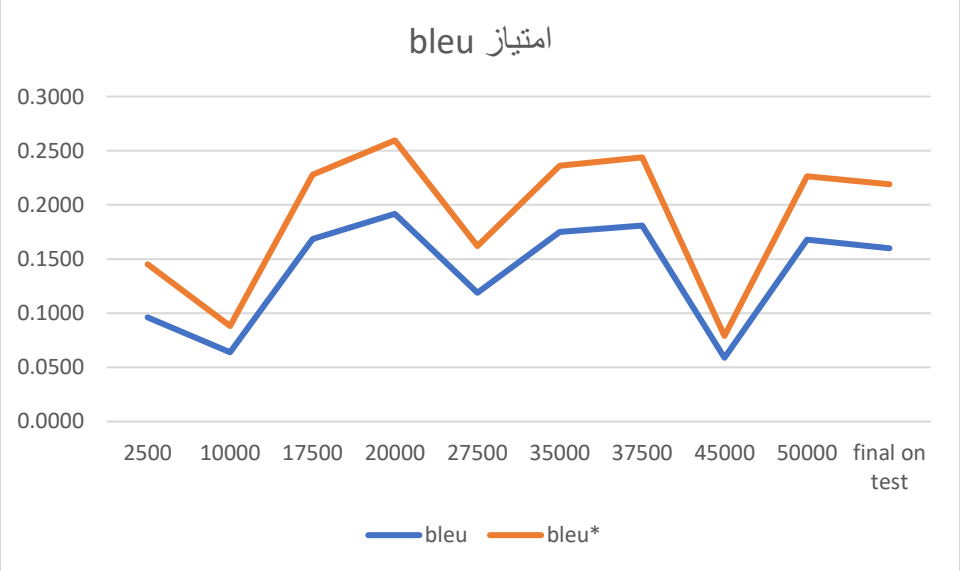
خروجی و نتایج

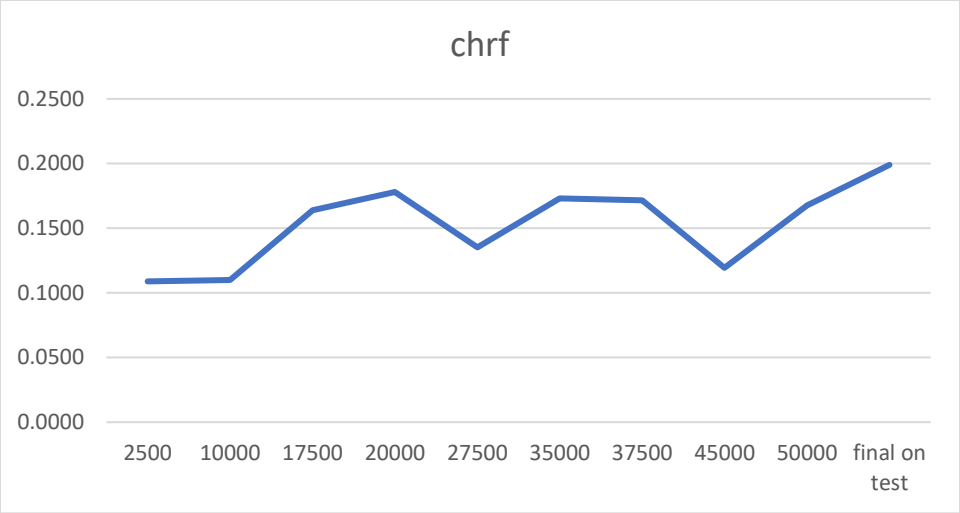
برای این بخش از کتابخانه nltk استفاده و معیار های bleu، chrf، gleu و nist برای فایل valid در گام‌های متفاوت و در نهایت بر روی داده‌گان تست محاسبه شده است.
نتایج مدل در چند گام مختلف با ارزیاب‌ها مختلف بررسی شده است

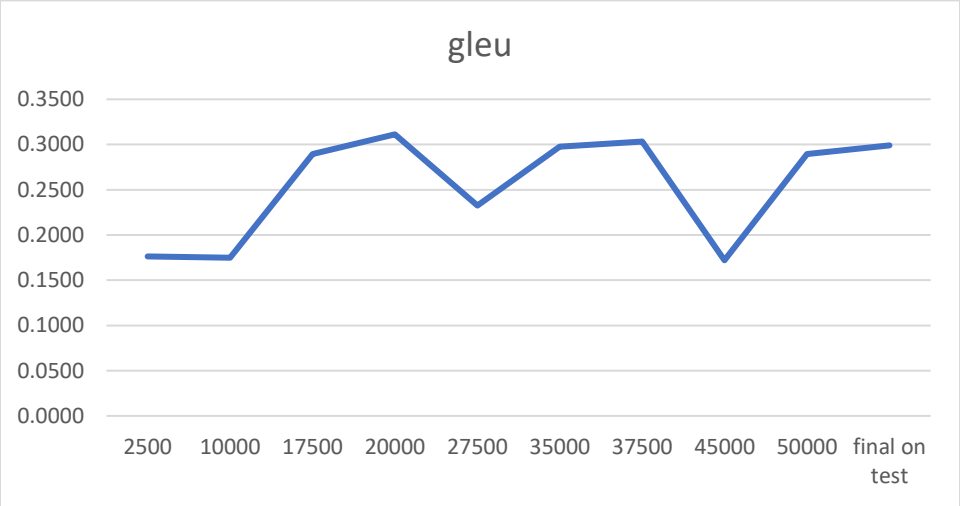
Table 1 امتیاز هر مدل با توجه به ارزیاب‌های متفاوت

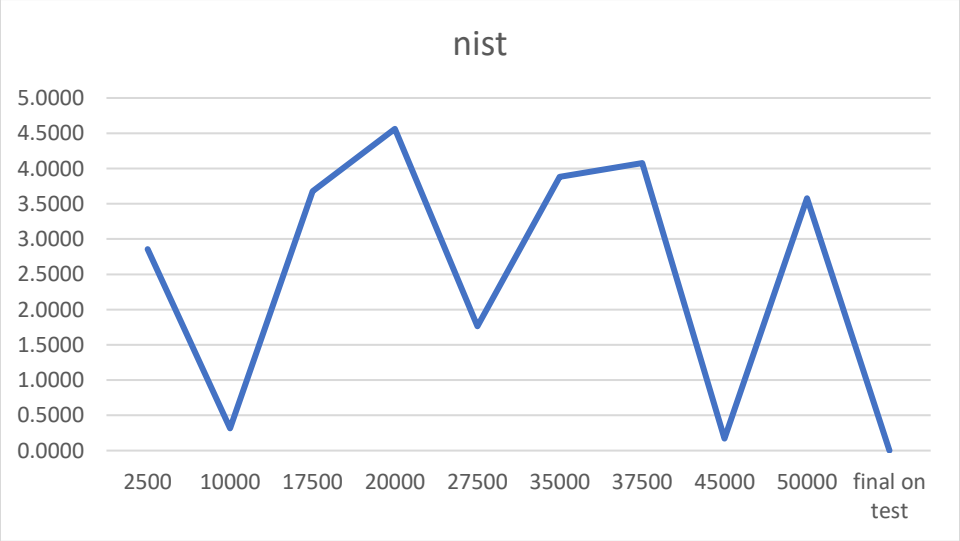
	2500	10000	17500	20000	27500	35000	37500	45000	50000	final on test
bleu	0.0961	0.0640	0.1686	0.1917	0.1188	0.1750	0.1809	0.0589	0.1680	0.1598
bleu*	0.1453	0.0882	0.2280	0.2596	0.1619	0.2360	0.2437	0.0790	0.2264	0.2189
chrf	0.1089	0.1100	0.1639	0.1780	0.1353	0.1731	0.1717	0.1193	0.1676	0.1989
gleu	0.1763	0.1747	0.2893	0.3113	0.2324	0.2977	0.3033	0.1720	0.2896	0.2990
nist	2.8551	0.3151	3.6763	4.5626	1.7618	3.8806	4.0782	0.1715	3.5801	3.2E-5

*نکته: در *Bleu وزن‌ها یکسان برای n-gram‌ها در نظر گرفته شده است که برابر با ۲۵٪ است ولی در bleu وزن‌ها متفاوت برای هر n-gram مشخص شده که این وزن‌ها برای 1-gram، 2-gram، 3-gram و 4-gram به‌ترتیب ۱٪، ۲٪، ۳٪ و ۴٪ است.*




chrf
0.2500
0.2000
0.1500
0.1000
0.0500
0.0000
2500
10000
17500
20000
27500
35000
37500
45000
50000
final on test


gleu
0.3500
0.3000
0.2500
0.2000
0.1500
0.1000
0.0500
0.0000
2500
10000
17500
20000
27500
35000
37500
45000
50000
final on test


nist
5.0000
4.5000
4.0000
3.5000
3.0000
2.5000
2.0000
1.5000
1.0000
0.5000
0.0000
2500
10000
17500
20000
27500
35000
37500
45000
50000
final on test

ابزار Marian-NMT

در این بخش قصد داریم مدلی با عملکرد مشابه ابزار Open-NMT این بار با ابزار Marian-NMT ایجاد کنیم. ماریان یک فریمورک بهینه ترجمه ماشینی بر پایه شبکه‌های عصبی مصنوعی است که تنها به‌وسیله زبان C++ توسعه داده شده است و از سریع‌ترین ابزارهای ممکن برای ایجاد یک مدل ترجمه ماشینی است. این فریمورک از جدیدترین معماری‌های شبکه عصبی مانند RNN و Transformer بهره می‌برد و توان استفاده از پردازنده گرافیکی را برای تسریع عملکرد خود دارا می‌باشد.

مقادیر و کانفیگ

در ابتدا نیاز است که برای هر کدام از زبان‌های موجود در فرآیند ترجمه ماشینی یک دیکشنری ایجاد کنیم. این دیکشنری با استفاده از مجموعه داده‌های آموزش زبان فارسی و انگلیسی به‌طور جداگانه برای هر زبان انجام می‌شود. عمل ذکر شده به شکل زیر و با استفاده از پارامترهای مشخص شده در محیط رابط متنی سیستم عامل لینوکس انجام می‌شود.

```
!./marian-vocab < /content/train.en > en_vocab.yml
!./marian-vocab < /content/train.fa > fa_vocab.yml
```

با استفاده از دستورات فوق برای هر یک از زبان‌های فارسی و انگلیسی فایل مجموعه داده آموزش که به‌ترتیب با نام‌های train.fa و train.en مشخص شده‌اند به‌عنوان ورودی به دستور marian-vocab بین علائم < و > قرار می‌گیرند و دیکشنری ایجاد شده در فایل‌های fa_vocab.yml و en_vocab.yml ذخیره می‌شوند. این دیکشنری‌ها در کنار مجموعه داده‌های آموزش و اعتبارسنجی(Validation) برای آموزش مدل الزامی می‌باشند.

```
!./marian/build/marian \
  --train-sets train.en train.fa \
  --vocabs en_vocab.yml fa_vocab.yml \
  --valid-sets valid.en valid.fa \
  --valid-script-path validate.sh \
  --valid-metrics bleu \
  --model /content/drive/MyDrive/NLP_HW5_data/Model/model.npz \
```

در مرحله بعد می‌توان مدل را با استفاده از داده‌های معرفی شده و دستور زیر آموزش داد:

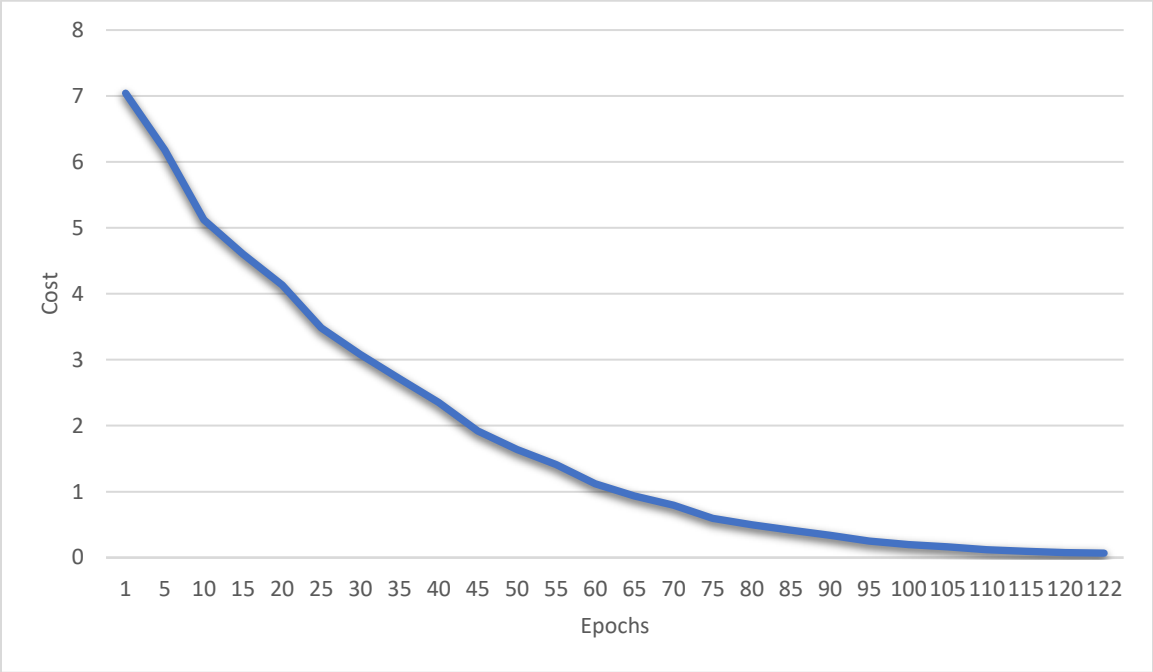
در ادامه تعدادی از پارامترهای این مدل معرفی می‌شوند. برخی از این پارامترها توسط ما مقداردهی شدند و برخی دیگر دارای مقدار پیش فرض مناسبی بودند و بنابراین مقدار دهی مجدد نشدند. جزئیات تمامی پارامترها پس از اجرای دستور فوق در محیط رابط متنی قابل مشاهده است.

۱. Train-sets: این پارامتر آدرس مجموعه داده‌های دو زبان مورد نظر برای ایجاد مدل ترجمه ماشینی که در این مدل زبان‌های فارسی و انگلیسی هستند را به‌عنوان مقادیر مناسب می‌پذیرد.
۲. Vocabs: آدرس دیکشنری‌های ایجاد شده در مرحله قبل مقادیر صحیح برای این پارامتر هستند.
۳. Valid-sets: مقادیر داده شده به این پارامتر آدرس بخش اعتبارسنجی مجموعه داده‌های زبان‌های فارسی و انگلیسی می‌باشند.
۴. Valid-metrics: این پارامتر نام معیارهای ارزیابی مورد نظر برای آموزش مدل را می‌پذیرد و همان طور که در صورت تمرین خواسته شده است، معیار ارزیابی مورد نظر برای هر دو مدل ایجاد شده bleu می‌باشد.
۵. Model: این پارامتر محل ذخیره مدل را نشان می‌دهد.
۶. نمادهای کلاس، ماسک و جداسازی BERT: این پارامترها به‌صورت پیش‌فرض دارای مقادیر صحیح [CLS]، [MASK] و [SEP] می‌باشند.
۷. Dim-emb: این پارامتر که نشان‌دهنده اندازه ابعاد embedding است به‌طور پیش‌فرض دارای مقدار ۵۱۲ می‌باشد که مقدار مناسبی برای روش‌های مبتنی برای Transfomer مانند BERT است.
۸. Disp-freq: فرکانس نمایش جزئیات معیارهای ارزیابی مانند bleu را نشان می‌دهد و دارای مقدار ۱۰۰۰ است.

۹. Early-stopping: این پارامتر نشان می‌دهد که بعد از چند تکرار با تغییر هزینه (Cost) به میزان کمتر از آستانه مورد نظر، مدل اجازه پایان فرآیند آموزش را دارد. مقدار پیش‌فرض این پارامتر برابر ۱۰ می‌باشد.

۱۰. Save-freq: نشان‌دهنده فرکانس ذخیره checkpointها می‌باشد و دارای مقدار ۱۰۰۰۰ است.

پس از اجرای دستور فوق با پارامترهای معرفی شده و مقادیر مناسب برای آن‌ها مدل شروع به آموزش دیدن می‌کند. این مدل در حین فرآیند آموزش بعد از هر Epoch مقدار هزینه را نشان می‌دهد. مقادیر هزینه در طی این فرآیند به شکل زیر تغییر می‌کند:



همان‌گونه‌که مشخص است پس از انجام محاسبات در Epoch ۱۲۲ و صرف زمان تقریبی ۲ ساعت مقدار هزینه در حال همگرا شدن به مقدار صفر است و در نتیجه می‌توان بهبود نتایج را از این مدل انتظار داشت. همچنین پس از هر ۱۰۰۰۰ تکرار تعدادی از برترین جملات ترجمه شده نیز نمایش داده می‌شود که تعدادی از آن‌ها برای هر تکرار در جدول زیر نمایش داده می‌شود.

تکرار	رتبه	جمله
۱۰۰۰۰	۱	این کار ممکن است که این کار را تایید کند
	۲	هیچ‌وقت به نظر نمی‌رسید
	۳	نمی‌دانم این کار را از دست می‌دادم
۲۰۰۰۰	۱	این چیزی که از آنچه ندیده شده‌است گفته بود
	۲	انگار خیلی نگران شده بود
	۴	این گزارش این است که سه ساعت بعد از ۳۳ امتیاز به‌عنوان ۲ میلیون متر می‌باشد
۳۰۰۰۰	۱	معلوم است که این گونه چیزی در پیش دیده می‌شد
	۲	همین قدر بیش‌تر به دنبالش نرسد
	۳	می دونم که دلم می‌خواهد این موضوع را باور کنم
۴۰۰۰۰	۱	اندیشه آن‌ها است که این بار دیگر در انتظارش بود
	۲	اکنون چنان به نظرشان خوش نداشت
	۳	من نمی‌دانم

۵۰۰۰۰	۱	این کلمه آن‌قدر است که چه فکر کند که
	۲	اکنون حالتی خراب کردن برود
	۳	من خوشحالم که نمی‌دانم چه چنین کاری را انکار کنم

آنچه از خروجی‌های مدل در هر مرحله می‌توان متوجه شد این است که نه تنها که مدل نسبت به مقدار Cost پیشرفت داشته است بلکه نتایج ترجمه داده‌های اعتبارسنجی نیز در هر مرحله بهتر شده‌اند. اگرچه به نظر می‌رسد که کیفیت برخی جملات مراحل آخر نسبت به مراحل قبلی پسرفت داشته است، باید این نکته را بدانیم که تنها فاکتور مهم در ترجمه ماشینی دقت ترجمه نیست و توانمندی این سیستم در ترجمه جملاتی با طول و لغات و سطح پیچیدگی مشخص می‌شود. برای مثال پس از ۱۰۰۰۰ تکرار جملاتی با طول بسیار کم و لغات بسیار معمول و پرکاربرد بهترین نتایج مدل هستند و نشان‌دهنده این است که مدل در ترجمه جملات طولانی‌تر و پیچیده‌تر ناتوان بوده است و این جملات امتیازات کمتری از معیارهای ارزیابی دریافت کرده‌اند. برخلاف مراحل اول، پس از ۵۰۰۰۰ تکرار جملاتی طولانی‌تر و با پیچیدگی بالاتر ترجمه شده‌اند. برای مثال جمله ترجمه شده "من خوشحالم که نمی‌دانم چه چنین کاری را انکار کنم" اگرچه دارای ایرادات جزئی می‌باشد و می‌توانست بهتر از این باشد اما دارای ساختار مناسبی است. نتیجه‌گیری فوق نشان می‌دهد که عملکرد مدل در طول مدت آموزش کدام در حال بهتر شدن بوده است و در صورت آموزش مدل در زمانی مناسب می‌توان به دقت مناسبی در ترجمه ماشینی با مجموعه داده موجود در تمرین دست یافت.